

# CENTRAL LIMIT THEOREMS AND MULTIPLIER BOOTSTRAP WHEN $p$ IS MUCH LARGER THAN $n$

VICTOR CHERNOZHUKOV, DENIS CHETVERIKOV, AND KENGO KATO

**ABSTRACT.** We derive a central limit theorem for the maximum of a sum of high dimensional random vectors. Specifically, we establish conditions under which the distribution of the maximum is approximated by that of the maximum of a sum of the Gaussian random vectors with the same covariance matrices as the original vectors. The key innovation of this result is that it applies even when the dimension of random vectors ( $p$ ) is large compared to the sample size ( $n$ ); in fact,  $p$  can be much larger than  $n$ . We also show that the distribution of the maximum of a sum of the random vectors with unknown covariance matrices can be consistently estimated by the distribution of the maximum of a sum of the conditional Gaussian random vectors obtained by multiplying the original vectors with i.i.d. Gaussian multipliers. This is the multiplier bootstrap procedure. Here too,  $p$  can be large or even much larger than  $n$ . These distributional approximations, either Gaussian or conditional Gaussian, yield a high-quality approximation to the distribution of the original maximum, often with approximation error decreasing polynomially in the sample size, and hence are of interest in many applications. We demonstrate how our central limit theorem and the multiplier bootstrap can be used for high dimensional estimation, multiple hypothesis testing, and adaptive specification testing. All these results contain non-asymptotic bounds on approximation errors.

## 1. INTRODUCTION

Let  $x_1, \dots, x_n$  be independent random vectors in  $\mathbb{R}^p$ , with each  $x_i$  having coordinates denoted by  $x_{ij}$ , i.e.,  $x_i = (x_{i1}, \dots, x_{ip})'$ . Suppose that each  $x_i$  is centered, namely  $E[x_i] = 0$ , and has a finite covariance matrix  $E[x_i x_i']$ . Consider the rescaled average:

$$(1) \quad X := (X_1, \dots, X_p)' := \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i.$$

---

*Date:* First version: June 2012. First Arxiv version: December, 2012. This version: January 25, 2013.

*Key words and phrases.* Dantzig selector, Slepian, Stein method, maximum of vector sums, high dimensionality, anti-concentration.

V. Chernozhukov and D. Chetverikov are supported by a National Science Foundation grant. K. Kato is supported by the Grant-in-Aid for Young Scientists (B) (22730179), the Japan Society for the Promotion of Science.

Our goal is to obtain a distributional approximation for the statistic  $T_0$  defined as the maximum coordinate of vector  $X$ :

$$T_0 := \max_{1 \leq j \leq p} X_j,$$

The distribution of  $T_0$  is of interest in many applications. When  $p$  is fixed, this distribution can be approximated by the classical Central Limit Theorem (CLT) applied to  $X$ . However, in modern applications, cf. [10],  $p$  is often comparable or even larger than  $n$ , and the classical CLT does not apply in such cases. This paper provides a tractable approximation to the distribution of  $T_0$  when  $p$  is large and possibly much larger than  $n$ .

The *first* main result of the paper is the Gaussian approximation theorem, which bounds the Kolmogorov distance between the distributions of  $T_0$  and its Gaussian analog  $Z_0$ . Specifically, let  $y_1, \dots, y_n$  be independent centered Gaussian random vectors in  $\mathbb{R}^p$  such that each  $y_i$  has the same covariance matrix as  $x_i$ , namely  $y_i \sim N(0, E[x_i x_i'])$ . Consider the rescaled average of these vectors,

$$(2) \quad Y := (Y_1, \dots, Y_p)' := \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i.$$

Vector  $Y$  is the Gaussian analog of  $X$  in the sense of sharing the same mean and covariance matrix, namely  $E[X] = E[Y] = 0$  and  $E[XX'] = E[YY'] = n^{-1} \sum_{i=1}^n E[x_i x_i']$ . We then define the Gaussian analog  $Z_0$  of  $T_0$  as the maximum coordinate of vector  $Y$ :

$$(3) \quad Z_0 := \max_{1 \leq j \leq p} Y_j.$$

Our main result shows that, under suitable moment assumptions, as  $n \rightarrow \infty$  and possibly  $p = p_n \rightarrow \infty$ ,

$$(4) \quad \rho := \sup_{t \in \mathbb{R}} |P(T_0 \leq t) - P(Z_0 \leq t)| \leq C n^{-c} \rightarrow 0,$$

where constants  $c > 0$  and  $C > 0$  are independent of  $n$ .

Importantly, in (4),  $p$  can be large in comparison to  $n$  and be nearly as large as  $e^{o(n^{1/7})}$ . For example, if  $x_{ij}$  are uniformly bounded (namely,  $|x_{ij}| \leq C_1$  for some constant  $C_1 > 0$  for all  $i$  and  $j$ ) the Kolmogorov distance  $\rho$  converges to zero at a polynomial rate whenever  $(\log p)^7/n \rightarrow 0$  at a polynomial rate. We obtain similar results when  $x_{ij}$  are sub-exponential and even non-sub-exponential under suitable moment assumptions. Figure 1 illustrates the result (4) in a non-subexponential example, which is motivated by the analysis of the Dantzig Selector of [11] in non-Gaussian settings (see Section 4).

The proof of the Gaussian approximation result (4) builds on a number of technical tools such as Slepian's smart path interpolation (which is related to the solution of Stein's partial differential equation; cf. Appendix E), Stein's leave-one-out method, approximation of maxima by the smooth

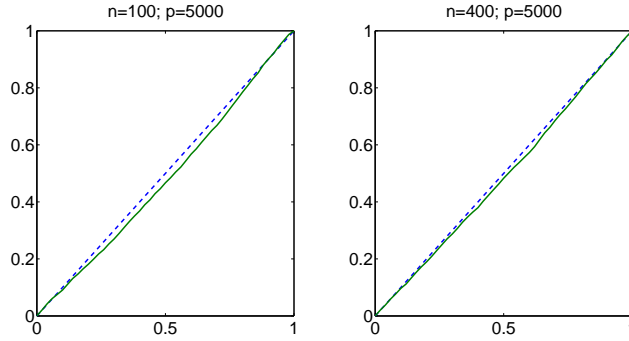


FIGURE 1. P-P plots comparing distributions of  $T_0$  and  $Z_0$  in the example motivated by the problem of selecting the penalty level of the Dantzig selector. Here  $x_{ij}$  are generated as  $x_{ij} = z_{ij}\varepsilon_i$  with  $\varepsilon_i \sim t(4)$ , (a  $t$ -distribution with four degrees of freedom), and  $z_{ij}$  are non-stochastic (simulated once using  $U[0, 1]$  distribution independently across  $i$  and  $j$ ). The dashed line is  $45^\circ$ . The distributions of  $T_0$  and  $Z_0$  are close, as (qualitatively) predicted by the CLT derived in the paper: see Corollaries 2.1 or 2.2. The quality of the Gaussian approximation is particularly good for the tail probabilities, which is most relevant for practical applications.

functions (related to “free energy” in spin glasses), and exponential inequalities for self-normalized sums. See, e.g., [41, 42, 20, 15, 43, 12, 19, 36] for introduction and discussion of some of these tools. It also critically relies on the anti-concentration and comparison bounds of maxima of Gaussian vectors derived in [18] and restated in this paper as Lemmas 2.1 and 3.1.

Our new Gaussian approximation theorem has the following innovative features. To the best of our knowledge, this is the first general result that establishes that maxima of sums of random vectors can be approximated in distribution by the maxima of sums of Gaussian random vectors when  $p \gg n$  and especially when  $p$  is of order  $e^{n^c}$  for some  $c > 0$ . The existing techniques can also lead to results of the form (4) when  $p = p_n \rightarrow \infty$ , but under much stronger conditions on  $p$ . For example, Yurinskii’s coupling implies (4) but requires  $p^5/n \rightarrow 0$ ; see Example 17 (Section 10) in [37]. Second, our Gaussian approximation theorem covers cases where  $T_0$  does not have a limit distribution as  $n \rightarrow \infty$  and  $p = p_n \rightarrow \infty$ . In some cases, after a suitable normalization,  $T_0$  could have an extreme value distribution as a limit distribution, but the approximation to an extreme value distribution requires some restrictions on the dependency structure among the coordinates in  $x_i$ . Our result does not require such restrictions on the dependency structure. Third, the quality of approximation in (4) is of polynomial order in  $n$ , which is better than the logarithmic in  $n$  quality that we could obtain in some (though not all) applications using the approximation of the distribution of  $T_0$  by an extreme value distribution (see [33]).

Our result also contributes to the literature on multivariate central limit theorems, which are concerned with conditions under which

$$(5) \quad |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| \rightarrow 0,$$

uniformly in a collection of sets  $A$ , typically *all* convex sets. Such results were developed among others, by [35, 38, 25, 7, 14], under conditions of type  $p^c/n \rightarrow 0$  (also see [13]). These results rely on the anti-concentration results for Gaussian random vectors on the  $\delta$ -expansions of boundaries of arbitrary convex sets  $A$  (see [4]). Note that our result also establishes (5), but uniformly for all convex sets of the form  $A_{\max} = \{a \in \mathbb{R}^p : \max_{1 \leq j \leq p} a_j \leq t\}$  for  $t \in \mathbb{R}$ . These sets have a rather special structure that allows us to deal with  $p \gg n$ : in particular, concentration of measure on the  $\delta$ -expansion of boundary of  $A_{\max}$  is at most of order  $\delta\sqrt{\log p}$  for Gaussian random vectors with unit variance, as shown in [18] (see also Lemma 2.1). (The relation (5) with  $A = A_{\max}$  explains the sense in which we have a CLT, as appearing in the title of the paper.)

Note that the result (4) is immediately useful for inference with statistic  $T_0$ , even though  $\mathbb{P}(Z_0 \leq t)$  needs not converge itself to a well behaved distribution function. Indeed, if the covariance matrix  $n^{-1} \sum_{i=1}^n \mathbb{E}[x_i x_i']$  is known, then  $c_{Z_0}(1 - \alpha) := (1 - \alpha)$ -quantile of  $Z_0$ , can be computed numerically, and we have

$$(6) \quad |\mathbb{P}(T_0 \leq c_{Z_0}(1 - \alpha)) - (1 - \alpha)| \leq Cn^{-c} \rightarrow 0.$$

A chief application of this kind arises in determination of the penalty level for the Dantzig selector of [11] in the high-dimensional regression with non-Gaussian errors, which we examine in Section 5. There, under the canonical (homoscedastic) noise, the covariance matrix is known, and so quantiles of  $Z_0$  can be easily computed numerically and used for choosing the penalty level. However, if the noise is heteroscedastic, the covariance matrix is no longer known, and this approach is no longer feasible. This motivates our second main result.

The *second* main result of the paper establishes validity of the multiplier bootstrap for estimating quantiles of  $Z_0$  when the covariance matrix  $n^{-1} \sum_{i=1}^n \mathbb{E}[x_i x_i']$  is unknown. More precisely, we define the Gaussian-symmetrized version  $W_0$  of  $T_0$  by multiplying  $x_i$  with i.i.d. standard Gaussian random variables  $e_1, \dots, e_n$ :

$$(7) \quad W_0 := \max_{1 \leq j \leq p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} e_i.$$

We show that the conditional quantiles of  $W_0$  given data  $(x_i)_{i=1}^n$  are able to consistently estimate the quantiles of  $Z_0$  and hence those of  $T_0$  (where the notion of consistency used is the one that guarantees asymptotically valid inference). Here the primary factor driving the bootstrap estimation error is the maximum difference between the empirical and population covariance

matrices:

$$\Delta := \max_{1 \leq j, k \leq p} \left| \frac{1}{n} \sum_{i=1}^n (x_{ij}x_{ik} - \mathbb{E}[x_{ij}x_{ik}]) \right|,$$

which can converge to zero even when  $p$  is much larger than  $n$ . For example, when  $x_{ij}$  are uniformly bounded, the multiplier bootstrap is valid for inference if  $(\log p)^7/n \rightarrow 0$ . Earlier related results on bootstrap in the “ $p \rightarrow \infty$  but  $p/n \rightarrow 0$ ” regime were obtained in [34]; interesting results for the case  $p \gg n$  based on concentration inequalities and symmetrization are studied in [2, 3], albeit the approach and results are quite different from those given here. In particular, in [2], either Gaussianity or symmetry in distribution is imposed on the data.

The key motivating example of our analysis is the high-dimensional sparse regression model. In this model, [11] and [8] assume Gaussian errors to analyze the Dantzig selector and Lasso. Our results show that Gaussianity is not necessary and the Gaussian-like conclusions hold approximately, with just the fourth moment of the regression errors being bounded. Moreover, our approximation allows to take into account correlations among the regressors. This leads to a better choice of the penalty level and tighter bounds on performance than those that had been available previously. For example, some of the same goals had been accomplished using moderate deviations for self-normalized sums, combined with the union bound [6]. However, the union bound does not take into account correlations among the regressors, and so it may be overly conservative in some applications.

Our results have a broad range of other applications. In addition to the high-dimensional estimation example, we show in the Supplemental Material how to apply our results in the multiple hypothesis testing via the step-down method of [40] and to specification testing. In either case number of hypotheses to be tested or the number of moment restrictions to be tested can be much larger than the sample size. Lastly, in a companion work ([17]), we are exploring the strong coupling for suprema of general empirical processes based on the methods developed here and maximal inequalities. These results represent a useful complement to the results based on the Hungarian coupling developed by [32, 9, 30, 39] for the entire empirical process and have applications to inference in nonparametric problems such as construction of uniform confidence bands (see, e.g., [24]).

**1.1. Organization of the paper.** In Section 2, we give the results on Gaussian approximation, and in Section 3 on the multiplier bootstrap. In Section 4, we present an application to the Dantzig selector. Appendices A-D contain proofs for each of these sections, with Appendix A stating auxiliary tools and lemmas. Due to the space limitation, we put additional results and proofs into Supplemental Material, Appendices G-F. In particular, Appendices G and H provide additional applications to multiple hypothesis and adaptive specification testing.

**1.2. Notation.** In what follows, unless otherwise stated, we will assume that  $p \geq 3$ . In making asymptotic statements we assume that  $n \rightarrow \infty$  with understanding that  $p$  depends on  $n$  and possibly  $p \rightarrow \infty$  as  $n \rightarrow \infty$ . Constants  $c, C, c_1, C_1, c_2, C_2, \dots$  are understood to be independent of  $n$ . Throughout the paper,  $\mathbb{E}_n[\cdot]$  denotes the average over index  $1 \leq i \leq n$ , i.e., it simply abbreviates the notation  $n^{-1} \sum_{i=1}^n [\cdot]$ . E.g.,  $\mathbb{E}_n[x_{ij}^2] = n^{-1} \sum_{i=1}^n x_{ij}^2$ . In addition,  $\bar{\mathbb{E}}[\cdot] = \mathbb{E}_n[\mathbb{E}[\cdot]]$ . For example,  $\bar{\mathbb{E}}[x_{ij}^2] = n^{-1} \sum_{i=1}^n \mathbb{E}[x_{ij}^2]$ . For a function  $f : \mathbb{R} \rightarrow \mathbb{R}$ , we write  $\partial^k f(x) = \partial^k f(x)/\partial x^k$  for nonnegative integer  $k$ ; for a function  $f : \mathbb{R}^p \rightarrow \mathbb{R}$ , we write  $\partial_j f(x) = \partial f(x)/\partial x_j$  for  $j = 1, \dots, p$ , where  $x = (x_1, \dots, x_p)'$ . Denote by  $C^k(\mathbb{R})$  the class of  $k$  times continuously differentiable functions from  $\mathbb{R}$  to itself, and denote by  $C_b^k(\mathbb{R})$  the class of all functions  $f \in C^k(\mathbb{R})$  such that  $\sup_{z \in \mathbb{R}} |\partial^j f(z)| < \infty$  for  $j = 0, \dots, k$ . We write  $a \lesssim b$  if  $a$  is smaller than or equal to  $b$  up to a universal positive constant. For  $a, b \in \mathbb{R}$ , we write  $a \vee b = \max\{a, b\}$ .

## 2. CENTRAL LIMIT THEOREMS FOR MAXIMA OF NON-GAUSSIAN SUMS

**2.1. Comparison Theorems and Non-Asymptotic Gaussian Approximations.** The purpose of this section is to compare and bound the difference between the expectations and distribution functions of the non-Gaussian to Gaussian maxima:

$$T_0 := \max_{1 \leq j \leq p} X_j \quad \text{and} \quad Z_0 := \max_{1 \leq j \leq p} Y_j,$$

where vector  $X$  is defined in equation (1) and  $Y$  in equation (2). Here and in what follows, without loss of generality, we will assume that  $(x_i)_{i=1}^n$  and  $(y_i)_{i=1}^n$  are independent. The following envelopes and bounds on moments will be used in stating the bounds in Gaussian approximations:

$$(8) \quad S_i := \max_{1 \leq j \leq p} (|x_{ij}| + |y_{ij}|), \quad M_k := \max_{1 \leq j \leq p} (\bar{\mathbb{E}}[x_{ij}^k])^{1/k}.$$

The problem of comparing distributions of maxima is of intrinsic difficulty since the maximum function  $z = (z_1, \dots, z_p)' \mapsto \max_{1 \leq j \leq p} z_j$  is non-differentiable. To circumvent the problem, we use a smooth approximation of the maximum function. For  $z = (z_1, \dots, z_p)' \in \mathbb{R}^p$ , consider the function:

$$F_\beta(z) := \beta^{-1} \log \left( \sum_{j=1}^p \exp(\beta z_j) \right),$$

which approximates the maximum function, where  $\beta > 0$  is the smoothing parameter that controls the level of approximation (we call this function the “smooth max function”). Indeed, an elementary calculation shows that for all  $z \in \mathbb{R}^p$ ,

$$(9) \quad 0 \leq F_\beta(z) - \max_{1 \leq j \leq p} z_j \leq \beta^{-1} \log p.$$

This smooth max function arises in the definition of “free energy” in spin glasses; see, e.g., [43].

We start with the following “warm-up” theorem that conveys the main qualitative feature of the problem. Here and in what follows, for a smooth function  $g : \mathbb{R} \rightarrow \mathbb{R}$ , write

$$G_k := \sup_{z \in \mathbb{R}} |\partial^k g(z)|, \quad k \geq 0.$$

**Theorem 2.1** (Comparison of Gaussian to Non-Gaussian Maxima, I). *For every  $g \in C_b^3(\mathbb{R})$  and  $\beta > 0$ ,*

$$|\mathbb{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathbb{E}}[S_i^3],$$

and hence

$$|\mathbb{E}[g(T_0) - g(Z_0)]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathbb{E}}[S_i^3] + \beta^{-1}G_1 \log p.$$

**Comment 2.1** (Optimizing the bound). The theorem bounds the difference between the expectations of smooth functions of maxima. The optimal value of the last bound is given by

$$\min_{\beta > 0} n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)\bar{\mathbb{E}}[S_i^3] + \beta^{-1}G_1 \log p.$$

We postpone choices of  $\beta$  to the proofs of subsequent corollaries, leaving ourselves more flexibility in optimizing bounds in those corollaries. ■

Deriving a bound on the Kolmogorov distance between distributions of  $T_0$  and  $Z_0$  from Theorem 2.1 is *not* a trivial issue and this step relies on the following *anti-concentration* inequality for maxima of Gaussian random variables, which is derived in [18].

**Lemma 2.1** (Anti-Concentration). *Let  $\xi_1, \dots, \xi_p$  be (not necessarily independent) centered Gaussian random variables with  $\sigma_j^2 := \mathbb{E}[\xi_j^2] > 0$  for all  $1 \leq j \leq p$ . Let  $\underline{\sigma} = \min_{1 \leq j \leq p} \sigma_j$  and  $\bar{\sigma} = \max_{1 \leq j \leq p} \sigma_j$ . Then for every  $\varsigma > 0$ ,*

$$\sup_{z \in \mathbb{R}} \mathbb{P} \left( \left| \max_{1 \leq j \leq p} \xi_j - z \right| \leq \varsigma \right) \leq C\varsigma \sqrt{1 \vee \log(p/\varsigma)},$$

where  $C > 0$  is a constant depending only on  $\underline{\sigma}$  and  $\bar{\sigma}$ . When  $\sigma_j$  are all equal,  $\log(p/\varsigma)$  on the right side can be replaced by  $\log p$ .

By Theorem 2.1 and Lemma 2.1, we can now derive a bound on the Kolmogorov distance between distributions of  $T_0$  and  $Z_0$ .

**Corollary 2.1** (Central Limit Theorem, I). *Suppose that there are some constants  $c_1 > 0$  and  $C_1 > 0$  such that  $c_1 \leq \bar{\mathbb{E}}[x_{ij}^2] \leq C_1$  for all  $1 \leq j \leq p$ . Then there exists a constant  $C > 0$  depending only on  $c_1$  and  $C_1$  such that*

$$\rho := \sup_{t \in \mathbb{R}} |\mathbb{P}(T_0 \leq t) - \mathbb{P}(Z_0 \leq t)| \leq C(n^{-1}(\log(pn))^7)^{1/8}(\bar{\mathbb{E}}[S_i^3])^{1/4}.$$

**Comment 2.2** (Main qualitative feature: logarithmic dependence on  $p$ ). Theorem 2.1 and Corollary 2.1 imply that the error of approximating the maximum coordinate in the sum of independent random vectors by its

Gaussian analogue depends on  $p$  (possibly) only through  $\log p$ . This is the main qualitative feature of all the results in this paper. Note also that the term  $\bar{\mathbb{E}}[S_i^3]$  implicitly encodes the complexity of the vectors, in particular it will reflect the correlation structure of vectors  $X$  and  $Y$ . However, both Theorem 2.1 and Corollary 2.1 and all subsequent results given below do not limit the dependence among the coordinates in  $x_i$ . ■

**Comment 2.3** (Motivation for the next result). While Theorem 2.1 and Corollary 2.1 convey an important qualitative aspect of the problem and admit easy-to-grasp proofs, an important disadvantage of these results is that the bounds depend on  $\bar{\mathbb{E}}[S_i^3]$ . If  $\bar{\mathbb{E}}[S_i^3] \leq C$ , Corollary 2.1 leads to  $\rho = O((n^{-1}(\log(pn))^7)^{1/8})$  and  $\rho \rightarrow 0$  as long as  $\log p = o(n^{1/7})$ . This is the case when, for example, as in caption to Figure 1,

$$x_{ij} = z_{ij}\varepsilon_i, \quad z_{ij} \text{ are non-stochastic with } |z_{ij}| \leq C, \quad \mathbb{E}[|\varepsilon_i|^3] \leq C.$$

When  $\bar{\mathbb{E}}[S_i^3]$  increases with  $n$ , however, the bounds need not be as good, and can be improved considerably by using a truncation method. Using such a method in conjunction with the proof strategy of Theorem 2.1, we derive in Theorem 2.2 below a bound that can be much better in the latter scenario. The improvement here comes at a cost of a more involved statement, involving truncation parameters. ■

To derive our next main result, we employ a truncation method. Given a threshold level  $u > 0$ , define a truncated version of  $x_{ij}$  by

$$(10) \quad \tilde{x}_{ij} = x_{ij}1\left\{|x_{ij}| \leq u(\bar{\mathbb{E}}[x_{ij}^2])^{1/2}\right\} - \mathbb{E}\left[x_{ij}1\left\{|x_{ij}| \leq u(\bar{\mathbb{E}}[x_{ij}^2])^{1/2}\right\}\right].$$

Let  $\varphi_x(u)$  be the infimum, which is attained, over all numbers  $\varphi \geq 0$  such that

$$(11) \quad \bar{\mathbb{E}}\left[x_{ij}^21\left\{|x_{ij}| > u(\bar{\mathbb{E}}[x_{ij}^2])^{1/2}\right\}\right] \leq \varphi^2\bar{\mathbb{E}}[x_{ij}^2].$$

Note that the function  $\varphi_x(u)$  is right-continuous; it measures the impact of truncation on second moments. Define  $u_x(\gamma)$  as the infimum  $u \geq 0$  such that

$$\mathbb{P}\left(|x_{ij}| \leq u(\bar{\mathbb{E}}[x_{ij}^2])^{1/2}, 1 \leq i \leq n, 1 \leq j \leq p\right) \geq 1 - \gamma.$$

Also define  $\varphi_y(u)$  and  $u_y(\gamma)$  by the corresponding quantities for the analogue Gaussian case, namely with  $(x_i)_{i=1}^n$  replaced by  $(y_i)_{i=1}^n$  in the above definitions. Throughout the paper we use the following quantities:

$$\varphi(u) := \varphi_x(u) \vee \varphi_y(u), \quad u(\gamma) := u_x(\gamma) \vee u_y(\gamma).$$

Here is the main theorem of this section. Recall the definition of  $M_k$  in (8).

**Theorem 2.2** (Comparison of Gaussian to Non-Gaussian Maxima, II). *Let  $\beta > 0, u > 0$  and  $\gamma \in (0, 1)$  be such that  $2\sqrt{2}uM_2\beta/\sqrt{n} \leq 1$  and  $u \geq u(\gamma)$ . Then for every  $g \in C_b^3(\mathbb{R})$ ,*

$$|\mathbb{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \lesssim D_n(g, \beta, u, \gamma),$$



and hence

$$|\mathbb{E}[g(T_0) - g(Z_0)]| \lesssim D_n(g, \beta, u, \gamma) + \beta^{-1} G_1 \log p,$$

where

$$\begin{aligned} D_n(g, \beta, u, \gamma) &:= n^{-1/2} (G_3 + G_2 \beta + G_1 \beta^2) M_3^3 + (G_2 + \beta G_1) M_2^2 \varphi(u) \\ &\quad + G_1 M_2 \varphi(u) \sqrt{\log(p/\gamma)} + G_0 \gamma. \end{aligned}$$

By Theorem 2.2 and Lemma 2.1, we can obtain a bound on the Kolmogorov distance between the distribution functions of  $T_0$  and  $Z_0$ .

**Corollary 2.2 (Central Limit Theorem, II).** *Suppose that there are some constants  $0 < c_1 < C_1$  such that  $c_1 \leq \bar{\mathbb{E}}[x_{ij}^2] \leq C_1$  for  $1 \leq j \leq p$ . Then for every  $\gamma \in (0, 1)$ ,*

$$\rho \leq C \left[ n^{-1/8} (M_3^{3/4} \vee M_4^{1/2}) (\log(pn/\gamma))^{7/8} + n^{-1/2} (\log(pn/\gamma))^{3/2} u(\gamma) + \gamma \right],$$

where  $C > 0$  is a constant that depends on  $c_1$  and  $C_1$  only.

In applications it is useful to bound the upper function  $u(\gamma)$ . Here is a simple and effective way of doing this. Let  $h : [0, \infty) \rightarrow [0, \infty)$  be a Young-Orlicz modulus, i.e., a convex and strictly increasing function with  $h(0) = 0$ . Denote by  $h^{-1}$  the inverse function of  $h$ . Standard examples include the power function  $h(v) = v^q$  with inverse  $h^{-1}(\gamma) = \gamma^{1/q}$  and the exponential function  $h(v) = \exp(v) - 1$  with inverse  $h^{-1}(\gamma) = \log(\gamma + 1)$ . These functions describe how many moments the random variables have, for example, a random variable  $\xi$  has finite  $q$ -th moment if  $\mathbb{E}[|\xi|^q] < \infty$ , and is sub-exponential if  $\mathbb{E}[\exp(|\xi|/C)] < \infty$  for some  $C > 0$ . We refer to [44], Chapter 2.2, for further details on Young-Orlicz moduli.

**Lemma 2.2** (Bounds on the upper function  $u(\gamma)$ ). *Let  $h : [0, \infty) \rightarrow [0, \infty)$  be a Young-Orlicz modulus, and let  $B > 0$  and  $D > 0$  be constants such that  $(\mathbb{E}[x_{ij}^2])^{1/2} \leq B$  for all  $1 \leq i \leq n, 1 \leq j \leq p$ , and  $\bar{\mathbb{E}}[h(\max_{1 \leq j \leq p} |x_{ij}|/D)] \leq 1$ . Then under the condition of Corollary 2.2,*

$$u(\gamma) \leq C \max\{Dh^{-1}(n/\gamma), B\sqrt{\log(pn/\gamma)}\},$$

where  $C > 0$  is a constant that depends on  $c_1$  and  $C_1$  only.

In applications, parameters  $B$  and  $D$  (with  $M_3$  and  $M_4$  as well) are allowed to increase with  $n$ . The size of these parameters and the choice of the Young-Orlicz modulus are case-specific.

**2.2. Examples of Applications.** The purpose of this subsection is to obtain bounds on  $\rho$  for various leading examples frequently encountered in applications. We are concerned with simple conditions under which  $\rho$  decays polynomially in  $n$ .

Let  $c_1 > 0, c_2 > 0, C_1 > 0$  be some constants, and let  $B_n \geq 1$  be a sequence of constants. We allow for the case where  $B_n \rightarrow \infty$  as  $n \rightarrow \infty$ . We shall first consider applications where one of the following conditions is satisfied *uniformly* in  $1 \leq i \leq n$  and  $1 \leq j \leq p$ :

- (E.1)  $\bar{E}[x_{ij}^2] \geq c_1$  and  $\bar{E}[S_i^3] \leq C_1$ ;  
 (E.2)  $\bar{E}[x_{ij}^2] \geq c_1$  and  $E[\exp(|x_{ij}|/C_1)] \leq 2$ ;  
 (E.3)  $c_1 \leq \bar{E}[x_{ij}^2] \leq C_1$  and  $|x_{ij}| \leq B_n$ .

**Comment 2.4.** Condition (E.1) is perhaps the simplest example in this paper; under this condition application of Corollary 2.1 is effective. A concrete example with condition (E.1) satisfied is the case where  $x_{ij} = z_{ij}\varepsilon_i$ ,  $z_{ij}$  are non-stochastic with  $|z_{ij}| \leq C$ , and  $E[|\varepsilon_i|^3] \leq C$ . Conditions (E.2)-(E.5) are more elaborate, intended to cover cases where moments of the envelopes  $S_i$  and higher order moments  $M_3$  and  $M_4$  increase with  $n$ . In these cases the use of Corollary 2.1 is not effective, and we shall use Corollary 2.2 instead. Condition (E.2) covers vectors  $x_i$  made up from sub-exponential random variables, including sub-Gaussian as a special case; this example is quite often used in high-dimensional statistics. Condition (E.3) covers variables that are bounded by  $B_n$ , which may increase with  $n$ ; many applications, after a suitable truncation, can be covered by it. ■

We shall also consider regression applications where one of the following conditions is satisfied *uniformly in*  $1 \leq i \leq n$  and  $1 \leq j \leq p$ :

- (E.4)  $x_{ij} = z_{ij}\varepsilon_{ij}$ , where  $z_{ij}$  are non-stochastic with  $|z_{ij}| \leq B_n$ ,  $E_n[z_{ij}^2] = 1$ , and  $E[\varepsilon_{ij}] = 0$ ,  $E[\varepsilon_{ij}^2] \geq c_1$ , and  $E[\exp(|\varepsilon_{ij}|/C_1)] \leq 2$ ; or  
 (E.5)  $x_{ij} = z_{ij}\varepsilon_{ij}$ , where  $z_{ij}$  are non-stochastic with  $|z_{ij}| \leq B_n$ ,  $E_n[z_{ij}^2] = 1$ , and  $E[\varepsilon_{ij}] = 0$ ,  $E[\varepsilon_{ij}^2] \geq c_1$ , and  $E[\max_{1 \leq j \leq p} \varepsilon_{ij}^4] \leq C_1$ .

**Comment 2.5.** The last two cases cover examples that arise in high-dimensional regression, e.g., [11], which we shall revisit later in the paper. Typically,  $\varepsilon_{ij}$  are independent of  $j$  (i.e.,  $\varepsilon_{ij} = \varepsilon_i$ ) and hence  $E[\max_{1 \leq j \leq p} \varepsilon_{ij}^4] \leq C_1$  in condition (E.5) reduces to  $E[\varepsilon_i^4] \leq C_1$  (we allow  $\varepsilon_{ij}$  dependent on  $j$  so that Corollary 2.3 covers the multiple hypothesis testing example in Appendix G). Interestingly, these examples are also connected to spin glasses, see e.g., [43] and [36] ( $z_{ij}$  can be interpreted as generalized products of “spins” and  $\varepsilon_i$  as their random “interactions”). ■

**Corollary 2.3 (Central Limit Theorem in Leading Examples).** *Suppose that one of the following conditions is satisfied: (i) condition (E.1) and  $(\log(pn))^7/n \leq C_1 n^{-c_2}$ ; (ii) condition (E.2) and  $(\log(pn))^7/n \leq C_1 n^{-c_2}$ ; (iii) condition (E.3) and  $B_n^2(\log(pn))^7/n \leq C_1 n^{-c_2}$ ; (vi) condition (E.4) and  $B_n^2(\log(pn))^7/n \leq C_1 n^{-c_2}$ ; or (v) condition (E.5) and  $B_n^4(\log(pn))^7/n \leq C_1 n^{-c_2}$ . Then there exist constants  $c > 0$  and  $C > 0$  depending only on  $c_1, c_2$  and  $C_1$  such that*

$$\rho \leq C n^{-c}.$$

**Comment 2.6.** Cases (ii)-(v) indeed follow relatively directly from Corollary 2.2 with help of Lemma 2.2. Moreover, from Lemma 2.2, it is routine to find other conditions that lead to the conclusion of Corollary 2.3. ■

## 3. MULTIPLIER BOOTSTRAP

**3.1. A Gaussian-to-Gaussian Comparison Theorem.** The proofs of the main results in this section rely on the following lemma. Let  $V$  and  $Y$  be centered Gaussian random vectors in  $\mathbb{R}^p$  with covariance matrices  $\Sigma^V$  and  $\Sigma^Y$ , respectively. The following lemma compares the distribution functions of  $\max_{1 \leq j \leq p} V_j$  and  $\max_{1 \leq j \leq p} Y_j$  in terms of  $p$  and

$$\Delta_0 := \max_{1 \leq j, k \leq p} |\Sigma_{jk}^V - \Sigma_{jk}^Y|.$$

**Lemma 3.1** (Comparison of Distributions of Gaussian Maxima). *Suppose that there are some constants  $0 < c_1 < C_1$  such that  $c_1 \leq \Sigma_{jj}^Y \leq C_1$  for all  $1 \leq j \leq p$ . Then there exists a constant  $C > 0$  depending only on  $c_1$  and  $C_1$  such that*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P} \left( \max_{1 \leq j \leq p} V_j \leq t \right) - \mathbb{P} \left( \max_{1 \leq j \leq p} Y_j \leq t \right) \right| \leq C \Delta_0^{1/3} (1 \vee \log(p/\Delta_0))^{2/3}.$$

**Comment 3.1.** The result is derived in [18], and extends that of [12] who gave an explicit error in Sudakov-Fernique comparison of expectations of maxima of Gaussian vectors. ■

**3.2. Multiplier Bootstrap Theorems.** Suppose that we have a dataset  $(x_i)_{i=1}^n$  consisting of  $n$  independent centered random vectors  $x_i$  in  $\mathbb{R}^p$ . In this section we are interested in approximating quantiles of

$$(12) \quad T_0 = \max_{1 \leq j \leq p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij}$$

using the multiplier bootstrap method. Specifically, let  $(e_i)_{i=1}^n$  be a sequence of i.i.d.  $N(0, 1)$  variables independent of  $(x_i)_{i=1}^n$ , and let

$$(13) \quad W_0 = \max_{1 \leq j \leq p} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_{ij} e_i.$$

Then we define the multiplier bootstrap estimator of the  $\alpha$ -quantile of  $T_0$  as the conditional  $\alpha$ -quantile of  $W_0$  given  $(x_i)_{i=1}^n$ , i.e.,

$$c_{W_0}(\alpha) := \inf\{t \in \mathbb{R} : \mathbb{P}_e(W_0 \leq t) \geq \alpha\},$$

where  $\mathbb{P}_e$  is the probability measure induced by the multiplier variables  $(e_i)_{i=1}^n$  holding  $(x_i)_{i=1}^n$  fixed (i.e.,  $\mathbb{P}_e(W_0 \leq t) = \mathbb{P}(W_0 \leq t \mid (x_i)_{i=1}^n)$ ). The multiplier bootstrap theorem below provides a non-asymptotic bound on the bootstrap estimation error:

$$|\mathbb{P}(T_0 \leq c_{W_0}(\alpha)) - \alpha|.$$

Before presenting the theorem, we first give a simple useful lemma that is helpful in the proof of the theorem and in power analysis in applications. Define

$$c_{Z_0}(\alpha) := \inf\{t \in \mathbb{R} : \mathbb{P}(Z_0 \leq t) \geq \alpha\},$$

where  $Z_0 = \max_{1 \leq j \leq p} \sum_{i=1}^n y_{ij} / \sqrt{n}$  and  $(y_i)_{i=1}^n$  is a sequence of independent  $N(0, E[x_i x_i'])$  vectors. Recall that

$$\Delta = \max_{1 \leq j, k \leq p} |\mathbb{E}_n[x_{ij} x_{ik}] - \bar{E}[x_{ij} x_{ik}]|.$$

**Lemma 3.2** (Comparison of Quantiles, I). *Suppose that there are some constants  $0 < c_1 < C_1$  such that  $c_1 \leq E[x_{ij}^2] \leq C_1$  for all  $1 \leq j \leq p$ . Then for every  $\alpha \in (0, 1)$ ,*

$$P(c_{W_0}(\alpha) \leq c_{Z_0}(\alpha + \pi(\vartheta))) \geq 1 - P(\Delta > \vartheta),$$

$$P(c_{Z_0}(\alpha) \leq c_{W_0}(\alpha + \pi(\vartheta))) \geq 1 - P(\Delta > \vartheta),$$

where, for  $C_2 > 0$  denoting a constant depending only on  $c_1$  and  $C_1$ ,

$$\pi(\vartheta) := C_2 \vartheta^{1/3} (1 \vee \log(p/\vartheta))^{2/3}.$$

Recall that  $\rho := \sup_{t \in \mathbb{R}} |P(T_0 \leq t) - P(Z_0 \leq t)|$ . We are now in position to state the main theorem of this section.

**Theorem 3.1 (Validity of Multiplier Bootstrap, I).** *Suppose that for some constants  $0 < c_1 < C_1$ , we have  $c_1 \leq \bar{E}[x_{ij}^2] \leq C_1$  for all  $1 \leq j \leq p$ . Then for any  $\vartheta > 0$ ,*

$$\sup_{\alpha \in (0,1)} |P(T_0 \leq c_{W_0}(\alpha)) - \alpha| \leq \rho + \pi(\vartheta) + P(\Delta > \vartheta).$$

Theorem 3.1 provides a useful result for the case where the statistics are maxima of exact averages. There are many applications, however, where the relevant statistics arise as maxima of approximate averages. The following result shows that the theorem continues to apply if the approximation error of the relevant statistic by a maximum of an exact average can be suitably controlled. Specifically, suppose that a statistic of interest, say  $T = T(x_1, \dots, x_n)$  which may not be of the form (12), can be approximated by  $T_0$  of the form (12), and that the multiplier bootstrap is performed on a statistic  $W = W(x_1, \dots, x_n, e_1, \dots, e_n)$ , which may be different from (13) but still can be approximated by  $W_0$  of the form (13).

We require the approximation to hold in the following sense: there exist  $\zeta_1 \geq 0$  and  $\zeta_2 \geq 0$ , depending on  $n$  (and typically  $\zeta_1 \rightarrow 0, \zeta_2 \rightarrow 0$  as  $n \rightarrow \infty$ ), such that

$$(14) \quad P(|T - T_0| > \zeta_1) < \zeta_2,$$

$$(15) \quad P(P_e(|W - W_0| > \zeta_1) > \zeta_2) < \zeta_2.$$

We use the  $\alpha$ -quantile of  $W = W(x_1, \dots, x_n, e_1, \dots, e_n)$ , computed conditional on  $(x_i)_{i=1}^n$ :

$$c_W(\alpha) := \inf\{t \in \mathbb{R} : P_e(W \leq t) \geq \alpha\},$$

as an estimate of the  $\alpha$ -quantile of  $T$ .

**Lemma 3.3** (Comparison of Quantiles, II). *Suppose that condition (15) is satisfied. Then for every  $\alpha \in (0, 1)$ ,*

$$\begin{aligned} P(c_W(\alpha) &\leq c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geq 1 - \zeta_2, \\ P(c_{W_0}(\alpha) &\leq c_W(\alpha + \zeta_2) + \zeta_1) \geq 1 - \zeta_2. \end{aligned}$$

The next result provides a bound on the bootstrap estimation error.

**Theorem 3.2 (Validity of Multiplier Bootstrap, II).** *Suppose that, for some constants  $0 < c_1 < C_1$ , we have  $c_1 \leq \bar{E}[x_{ij}^2] \leq C_1$  for all  $1 \leq j \leq p$ . Moreover, suppose that conditions (14) and (15) are satisfied. Then for any  $\vartheta > 0$ ,*

$$\sup_{\alpha \in (0,1)} |P(T \leq c_W(\alpha)) - \alpha| \leq \rho + \pi(\vartheta) + P(\Delta > \vartheta) + C_3 \zeta_1 \sqrt{1 \vee \log(p/\zeta_1)} + \zeta_2,$$

where  $\pi(\cdot)$  is defined in Lemma 3.2, and  $C_3 > 0$  depends only on  $c_1$  and  $C_1$ .

**3.3. Examples of Applications: Revisited.** Here we revisit the examples in Section 2.2 and see how the multiplier bootstrap works for these leading examples. Let, as before,  $c_1 > 0, c_2 > 0$  and  $C_1 > 0$  be some constants, and let  $B_n \geq 1$  be a sequence of constants. Recall conditions (E.2)-(E.5) in Section 2.2.

**Corollary 3.1 (Multiplier Bootstrap in Leading Examples).** *Suppose that conditions (14) and (15) hold with  $\zeta_1 \sqrt{\log p} + \zeta_2 \leq C_1 n^{-c_2}$ . Moreover, suppose that one of the following conditions is satisfied: (i) condition (E.2) and  $(\log(pn))^7/n \leq C_1 n^{-c_2}$ ; (ii) condition (E.3), and  $B_n^2 (\log(pn))^7/n \leq C_1 n^{-c_2}$ ; (iii) condition (E.4) and  $B_n^2 (\log(pn))^7/n \leq C_1 n^{-c_2}$ ; or (iv) condition (E.5) and  $B_n^4 (\log(pn))^7/n \leq C_1 n^{-c_2}$ . Then there exist constants  $c > 0$  and  $C > 0$  depending only on  $c_1, c_2$  and  $C_1$  such that*

$$\sup_{\alpha \in (0,1)} |P(T \leq c_W(\alpha)) - \alpha| \leq C n^{-c}.$$

**Comment 3.2.** This corollary shows that the multiplier bootstrap is valid with a polynomial rate of accuracy for the significance level under weak conditions. This is in contrast with the extremal theory of Gaussian processes that provides only a logarithmic rate of approximation (see, e.g., [33] and [27]). ■

#### 4. APPLICATION: DANTZIG SELECTOR IN THE NON-GAUSSIAN MODEL

The purpose of this section is to demonstrate the case with which the CLT and the multiplier bootstrap theorem given in Corollaries 2.3 and 3.1 can be applied in important problems, dealing with a high-dimensional inference and estimation. We consider the Dantzig selector previously studied in the path-breaking works of [11], [8], [45] in the Gaussian setting and of [31] in a sub-exponential setting. Here we consider the non-Gaussian case, where the errors have only four bounded moments, and derive the performance bounds

that are approximately as sharp as in the Gaussian model. We consider both homoscedastic and heteroscedastic models.

**4.1. Homoscedastic case.** Let  $(z_i, y_i)_{i=1}^n$  be a sample of independent observations where  $z_i \in \mathbb{R}^p$  is a non-stochastic vector of regressors. We consider the model

$$y_i = z_i' \beta + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0, \quad i = 1, \dots, n, \quad \mathbb{E}_n[z_{ij}^2] = 1, \quad j = 1, \dots, p,$$

where  $y_i$  is a random scalar dependent variable, and the regressors are normalized in such a way that  $\mathbb{E}_n[z_{ij}^2] = 1$ . Here we consider the homoscedastic case:

$$\mathbb{E}[\varepsilon_i^2] = \sigma^2, \quad i = 1, \dots, n,$$

where  $\sigma^2$  is assumed to be known (for simplicity). We allow  $p$  to be substantially larger than  $n$ . It is well known that a condition that gives a good performance for the Dantzig selector is that  $\beta$  is sparse, namely  $\|\beta\|_0 \leq s \ll n$  (although this assumption will not be invoked below explicitly).

The aim is to estimate the vector  $\beta$  in some semi-norms of interest:  $\|\cdot\|_I$ . For example, given an estimator  $\hat{\beta}$  the prediction semi-norm for  $\delta = \hat{\beta} - \beta$  is

$$\|\delta\|_{\text{pr}} = \sqrt{\mathbb{E}_n[(z_i' \delta)^2]},$$

or the  $j$ -th component seminorm for  $\delta$  is

$$\|\delta\|_{\text{jc}} = |\delta_j|,$$

and so on. The label  $I$  designates the name of a norm of interest.

The Dantzig selector is the estimator defined by

$$(16) \quad \hat{\beta} \in \arg \min_{b \in \mathbb{R}^p} \|b\|_{\ell_1} \quad \text{subject to} \quad \sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}(y_i - z_i' b)]| \leq \lambda,$$

where  $\|\beta\|_{\ell_1} = \sum_{j=1}^p |\beta_j|$  is the  $\ell_1$ -norm. An ideal choice of the penalty level  $\lambda$  is meant to ensure that

$$T_0 := \sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij} \varepsilon_i]| \leq \lambda$$

with a prescribed probability  $1 - \alpha$ . Hence we would like to set penalty level  $\lambda$  equal to

$$c_{T_0}(1 - \alpha) := (1 - \alpha)\text{-quantile of } T_0,$$

(note that  $z_i$  are treated as fixed). Indeed, this penalty would take into account the correlation amongst the regressors, thereby adapting the performance of the estimator to the design condition. We can approximate this quantity using the central limit theorems derived in Section 2. Specifically, let

$$Z_0 := \sigma \sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij} e_i]|,$$

where  $e_i$  are i.i.d.  $N(0, 1)$  random variables independent of the data. We then estimate  $c_{T_0}(1 - \alpha)$  by

$$c_{Z_0}(1 - \alpha) := (1 - \alpha)\text{-quantile of } Z_0.$$

Note that we can calculate  $c_{Z_0}(1-\alpha)$  numerically with any specified precision by the simulation. (In a Gaussian model, design-adaptive penalty level  $c_{Z_0}(1-\alpha)$  was proposed in [5], but its extension to non-Gaussian cases was not available up to now).

An alternative choice of the penalty level is given by

$$c_0(1-\alpha) := \sigma \Phi^{-1}(1-\alpha/(2p)),$$

which is the canonical choice; see [11] and [8]. Note that canonical choice  $c_0(1-\alpha)$  disregards the correlation amongst the regressors, and is therefore more conservative than  $c_{Z_0}(1-\alpha)$ . Indeed, by the union bound, we see that

$$c_{Z_0}(1-\alpha) \leq c_0(1-\alpha).$$

Our first result below shows that the *either* of the two penalty choices,  $\lambda = c_{Z_0}(1-\alpha)$  or  $\lambda = c_0(1-\alpha)$ , are approximately valid under non-Gaussian noise—under the mild moment assumption  $E[\varepsilon_i^4] \leq \text{const.}$  replacing the canonical Gaussian noise assumption. To derive this result we apply our CLT to  $T_0$  to establish that the difference between distribution functions of  $T_0$  and  $Z_0$  approaches zero at polynomial speed. Indeed  $T_0$  can be represented as a maximum of averages,  $T_0 = \max_{1 \leq k \leq 2p} n^{-1/2} \sum_{i=1}^n \tilde{z}_{ik} \varepsilon_i$ , for  $\tilde{z}_i = (z'_i, -z'_i)'$ , and therefore our CLT applies.

To derive the bound on estimation error  $\|\delta\|_I$  in a seminorm of interest, we employ the following identifiability factor:

$$\kappa_I(\beta) := \inf_{\delta \in \mathbb{R}^p} \left\{ \max_{1 \leq j \leq p} \frac{|\mathbb{E}_n[z_{ij}(z'_i \delta)]|}{\|\delta\|_I} : \delta \in \mathcal{R}(\beta), \|\delta\|_I \neq 0 \right\},$$

where  $\mathcal{R}(\beta) := \{\delta \in \mathbb{R}^p : \|\beta + \delta\|_{\ell_1} \leq \|\beta\|_{\ell_1}\}$  is the restricted set;  $\kappa_I(\beta)$  is defined as  $\infty$  if  $\mathcal{R}(\beta) = \{0\}$  (this happens if  $\beta = 0$ ). The factors summarize the impact of sparsity of true parameter value  $\beta$  and the design on the identifiability of  $\beta$  with respect to the norm  $\|\cdot\|_I$ .

**Comment 4.1** (A comment on the identifiability factor  $\kappa_I(\beta)$ ). The identifiability factors  $\kappa_I(\beta)$  depend on the true parameter value  $\beta$ . This is not the main focus of this section, but we note that these factors represent a modest generalization of the cone invertibility factors and sensitivity characteristics defined in [45] and [23], which are known to be quite general. The main difference perhaps is the use of a norm of interest  $\|\cdot\|_I$  instead of the  $\ell_q$  norms and the use of smaller (non-conic) restricted set  $\mathcal{R}(\beta)$  in the definition. It is useful to note for later comparisons that in the case of prediction norm  $\|\cdot\|_I = \|\cdot\|_{\text{pr}}$  and under the exact sparsity assumption  $\|\beta\|_0 \leq s$ , we have

$$(17) \quad \kappa_{\text{pr}}(\beta) \geq 2^{-1} s^{-1/2} \kappa(s, 1),$$

where  $\kappa(s, 1)$  is the restricted eigenvalue defined in [8]. ■

Next we state bounds on the estimation error for the Dantzig selector  $\hat{\beta}^{(0)}$  with canonical penalty level  $\lambda = \lambda^{(0)} := c_0(1-\alpha)$  and the Dantzig selector  $\hat{\beta}^{(1)}$  with design-adaptive penalty level  $\lambda = \lambda^{(1)} := c_{Z_0}(1-\alpha)$ .

**Theorem 4.1** (Performance of Dantzig Selector in Non-Gaussian Model). *Suppose that there are some constants  $c_1 > 0, C_1 > 0$  and  $\sigma^2 > 0$ , and a sequence  $B_n \geq 1$  of constants such that for all  $1 \leq i \leq n$  and  $1 \leq j \leq p$ : (i)  $|z_{ij}| \leq B_n$ ; (ii)  $\mathbb{E}_n[z_{ij}^2] = 1$ ; (iii)  $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ ; (iv)  $\mathbb{E}[\varepsilon_i^4] \leq C_1$ ; and (v)  $B_n^4(\log(pn))^7/n \leq C_1 n^{-c_1}$ . Then there exist constants  $c > 0$  and  $C > 0$  depending only on  $c_1, C_1$  and  $\sigma^2$  such that, with probability at least  $1 - \alpha - Cn^{-c}$ , for either  $k = 0$  or 1,*

$$\|\hat{\beta}^{(k)} - \beta\|_I \leq \frac{2\lambda^{(k)}}{\sqrt{n\kappa_I(\beta)}}.$$

The most important feature of this result is that it provides Gaussian-like conclusions (as explained below) in a model with non-Gaussian noise, having only four bounded moments. However, the probabilistic guarantee is not  $1 - \alpha$  as, e.g., in [8], but rather  $1 - \alpha - Cn^{-c}$ , which reflects the cost of non-Gaussianity (along with more stringent side conditions). In what follows we discuss details of this result. Note that the bound above holds for any semi-norm of interest  $\|\cdot\|_I$ .

**Comment 4.2** (Improved Performance from Design-Adaptive Penalty Level). The use of the design-adaptive penalty level implies a better performance guarantee for  $\hat{\beta}^{(1)}$  over  $\hat{\beta}^{(0)}$ . Indeed, we have

$$\frac{2c_{Z_0}(1 - \alpha)}{\sqrt{n\kappa_I(\beta)}} \leq \frac{2c_0(1 - \alpha)}{\sqrt{n\kappa_I(\beta)}}.$$

E.g., in some designs, we can have  $\sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}e_i]| = O_P(1)$ , so that  $c_{Z_0}(1 - \alpha) = O(1)$ , whereas  $c_0(1 - \alpha) \propto \sqrt{\log p}$ . Thus, the performance guarantee provided by  $\hat{\beta}^{(1)}$  can be much better than that of  $\hat{\beta}^{(0)}$ . ■

**Comment 4.3** (Relation to the previous results under Gaussianity). To compare to the previous results obtained for the Gaussian settings, let us focus on the prediction norm and on estimator  $\hat{\beta}^{(1)}$  with penalty level  $\lambda = c_{Z_0}(1 - \alpha)$ . Suppose that the true value  $\beta$  is sparse, namely  $\|\beta\|_0 \leq s$ . In this case, with probability at least  $1 - \alpha - Cn^{-c}$ ,

$$(18) \quad \|\hat{\beta}^{(1)} - \beta\|_{\text{pr}} \leq \frac{2c_{Z_0}(1 - \alpha)}{\sqrt{n\kappa_{\text{pr}}(\beta)}} \leq \frac{4\sqrt{s}c_0(1 - \alpha)}{\sqrt{n\kappa(s, 1)}} \leq \frac{4\sqrt{s}\sqrt{2\log(\alpha/(2p))}}{\sqrt{n\kappa(s, 1)}},$$

where the last bound is the same as in [8], Theorem 7.1, obtained for the Gaussian case. We recover the same (or tighter) upper bound without making the Gaussianity assumption on the errors. However, the probabilistic guarantee is not  $1 - \alpha$  as in [8], but rather  $1 - \alpha - Cn^{-c}$ , which together with side conditions is the cost of non-Gaussianity. ■

**Comment 4.4** (Other refinements). Unrelated to the main theme of this paper, we can see from (18) that there is some tightening of the performance bound due to the use of the identifiability factor  $\kappa_{\text{pr}}(\beta)$  in place of the restricted eigenvalue  $\kappa(s, 1)$ ; for example, if  $p = 2$  and  $s = 1$  and the two regressors are identical, then  $\kappa_{\text{pr}}(\beta) > 0$ , whereas  $\kappa(1, 1) = 0$ . There is also



some tightening due to the use of  $c_{Z_0}(1 - \alpha)$  instead of  $c_0(1 - \alpha)$  as penalty level, as mentioned above.  $\blacksquare$

**4.2. Heteroscedastic case.** We consider the same model as above, except now the assumption on the error becomes

$$\sigma_i^2 := \mathbb{E}[\varepsilon_i^2] \leq \sigma^2, \quad i = 1, \dots, n,$$

i.e.,  $\sigma^2$  is the upper bound on the conditional variance, and we assume that this bound is known (for simplicity). As before, ideally we would like to set penalty level  $\lambda$  equal to

$$c_{T_0}(1 - \alpha) := (1 - \alpha)\text{-quantile of } T_0,$$

(where  $T_0$  is defined above, and we note that  $z_i$  are treated as fixed). The CLT applies as before, namely the difference of the distribution functions of  $T_0$  and its Gaussian analogue  $Z_0$  converges to zero. In this case, the Gaussian analogue can be represented as

$$Z_0 := \sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}\sigma_i e_i]|.$$

Unlike in the homoscedastic case, the covariance structure is no longer known, since  $\sigma_i$  are unknown and we can no longer calculate the quantiles of  $Z_0$ . However, we can estimate them using the following multiplier bootstrap procedure.

First, we estimate the residuals  $\hat{\varepsilon}_i = y_i - z_i' \hat{\beta}^{(0)}$  obtained from a preliminary Dantzig selector  $\hat{\beta}^{(0)}$  with the conservative penalty level  $\lambda = \lambda^{(0)} := c_0(1 - 1/n) := \sigma \Phi^{-1}(1 - 1/(2pn))$ , where  $\sigma^2$  is the upper bound on the error variance assumed to be known. Let  $(e_i)_{i=1}^n$  be a sequence of i.i.d. standard Gaussian random variables, and let

$$W := \sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}\hat{\varepsilon}_i e_i]|.$$

Then we estimate  $c_{Z_0}(1 - \alpha)$  by

$$c_W(1 - \alpha) := (1 - \alpha)\text{-quantile of } W,$$

defined conditional on data  $(z_i, y_i)_{i=1}^n$ . Note that  $c_W(1 - \alpha)$  can be calculated numerically with any specified precision by the simulation. Then we apply program (16) with  $\lambda = \lambda^{(1)} = c_W(1 - \alpha)$  to obtain  $\hat{\beta}^{(1)}$ .

**Theorem 4.2** (Performance of Dantzig in Non-Gaussian Model with Bootstrap Penalty Level). *Suppose that there are some constants  $c_1 > 0, C_1 > 0, \underline{\sigma}^2 > 0$  and  $\sigma^2 > 0$ , and a sequence  $B_n \geq 1$  of constants such that for all  $1 \leq i \leq n$  and  $1 \leq j \leq p$ : (i)  $|z_{ij}| \leq B_n$ ; (ii)  $\mathbb{E}_n[z_{ij}^2] = 1$ ; (iii)  $\underline{\sigma}^2 \leq \mathbb{E}[\varepsilon_i^2] \leq \sigma^2$ ; (iv)  $\mathbb{E}[\varepsilon_i^4] \leq C_1$ ; (v)  $B_n^4(\log(pn))^7/n \leq C_1 n^{-c_1}$ ; and (vi)  $(\log p)B_n c_0(1 - 1/n)/(\sqrt{n}\kappa_{\text{pr}}(\beta)) \leq C_1 n^{-c_1}$ . Then there exist constants  $c > 0$  and  $C > 0$  depending only on  $c_1, C_1, \underline{\sigma}^2$  and  $\sigma^2$  such that, with probability at least  $1 - \alpha - \nu_n$  where  $\nu_n = Cn^{-c}$ , we have*

$$(19) \quad \|\hat{\beta}^{(1)} - \beta\|_I \leq \frac{2\lambda^{(1)}}{\sqrt{n}\kappa_I(\beta)}.$$

Moreover, with probability at least  $1 - \nu_n$ ,

$$\lambda^{(1)} = c_W(1 - \alpha) \leq c_{Z_0}(1 - \alpha + \nu_n),$$

where  $c_{Z_0}(1 - a) := (1 - a)$ -quantile of  $Z_0$ ; in particular  $c_{Z_0}(1 - a) \leq c_0(1 - a)$ .

**4.3. Some Extensions.** Here we comment on some additional potential applications.

**Comment 4.5** (Confidence Sets). Note that bounds given in the preceding theorems can be used for inference on  $\beta$  or components of  $\beta$ , given the assumption  $\kappa_I(\beta) \geq \kappa$ , where  $\kappa$  is a known constant. For example, consider inference on the  $j$ -th component  $\beta_j$  of  $\beta$ . In this case, we take the norm of interest  $\|\delta\|_I$  to be  $\|\delta\|_{jc} = |\delta_j|$  on  $\mathbb{R}^p$ , and consider the corresponding identifiability factor  $\kappa_{jc}(\beta)$ . Suppose it is known that  $\kappa_{jc}(\beta) \geq \kappa$ . Then a  $(1 - \alpha - Cn^{-c})$ -confidence interval for  $\beta_j$  is given by

$$\{b \in \mathbb{R} : |\widehat{\beta}_j^{(1)} - b| \leq 2\lambda^{(1)} / (\sqrt{n}\kappa)\}.$$

This confidence set is of interest, but it does require the investigator to make a stance on what a plausible  $\kappa$  should be. We refer to [23] for a justification of confidence sets of this type and possible ways of computing lower bounds on  $\kappa$ ; there is also a work by [29], which provides computable lower bounds on related quantities. ■

**Comment 4.6** (Generalization of Dantzig Selector). There are many interesting applications where the results given above apply. There are, for example, interesting works by [1] and [22] that consider related estimators that minimize a convex penalty subject to the multiresolution screening constraints. In the context of the regression problem studied above, such estimators may be defined as:

$$\widehat{\beta} \in \arg \min_{b \in \mathbb{R}^p} J(b) \text{ subject to } \sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}(y_i - z'_i b)]| \leq \lambda,$$

where  $J$  is a convex penalty, and the constraint is used for multiresolution screening. For example, the Lasso estimator is nested by the above formulation by using  $J(b) = \|b\|_{\text{pr}}$ , and the previous Dantzig selector by using  $J(b) = \|b\|_{\ell_1}$ ; the estimators can be interpreted as a point in confidence set for  $\beta$ , which lies closest to zero under  $J$ -discrepancy (see references above for both of these points). Our results on choosing  $\lambda$  apply to this class of estimators, and the previous analysis also applies by re-defining the identifiability factor  $\kappa_I(\beta)$  relative to the new restricted set  $\mathcal{R}(\beta) := \{\delta \in \mathbb{R}^p : J(\beta + \delta) \leq J(\beta)\}$ ; where  $\kappa_I(\beta)$  is defined as  $\infty$  if  $\mathcal{R}(\beta) = \{0\}$ . ■

## APPENDIX A. PRELIMINARIES

**A.1. A Useful Maximal Inequality.** The following lemma, which is derived in [18], is a useful variation of standard maximal inequalities.

**Lemma A.1** (Maximal Inequality). *Let  $x_1, \dots, x_n$  be independent random vectors in  $\mathbb{R}^p$  with  $p \geq 2$ . Let  $M = \max_{1 \leq i \leq n} \max_{1 \leq j \leq p} |x_{ij}|$  and  $\sigma^2 = \max_{1 \leq j \leq p} \bar{\mathbb{E}}[x_{ij}^2]$ . Then*

$$\mathbb{E} \left[ \max_{1 \leq j \leq p} |\mathbb{E}_n[x_{ij}] - \bar{\mathbb{E}}[x_{ij}]| \right] \lesssim \sigma \sqrt{(\log p)/n} + \sqrt{\mathbb{E}[M^2]}(\log p)/n.$$

*Proof.* See [18], Lemma 8. ■

**A.2. Properties of the Smooth Max Function.** We will use the following properties of the smooth max function.

**Lemma A.2** (Properties of  $F_\beta$ ). *For every  $1 \leq j, k, l \leq p$ ,*

$$\partial_j F_\beta(z) = \pi_j(z), \quad \partial_j \partial_k F_\beta(z) = \beta w_{jk}(z), \quad \partial_j \partial_k \partial_l F_\beta(z) = \beta^2 q_{jkl}(z).$$

*where, for  $\delta_{jk} := 1\{j = k\}$ ,*

$$\begin{aligned} \pi_j(z) &:= e^{\beta z_j} / \sum_{m=1}^p e^{\beta z_m}, \quad w_{jk}(z) := (\pi_j \delta_{jk} - \pi_j \pi_k)(z), \\ q_{jkl}(z) &:= (\pi_j \delta_{jl} \delta_{jk} - \pi_j \pi_l \delta_{jk} - \pi_j \pi_k (\delta_{jl} + \delta_{kl}) + 2\pi_j \pi_k \pi_l)(z). \end{aligned}$$

*Moreover,*

$$\pi_j(z) \geq 0, \quad \sum_{j=1}^p \pi_j(z) = 1, \quad \sum_{j,k=1}^p |w_{jk}(z)| \leq 2, \quad \sum_{j,k,l=1}^p |q_{jkl}(z)| \leq 6.$$

*Proof of Lemma A.2.* The first property was noted in [12]. The other properties follow from repeated application of the chain rule. ■

**Lemma A.3** (Lipschitz Property of  $F_\beta$ ). *For every  $x \in \mathbb{R}^p$  and  $z \in \mathbb{R}^p$ , we have  $|F_\beta(x) - F_\beta(z)| \leq \max_{1 \leq j \leq p} |x_j - z_j|$ .*

*Proof of Lemma A.3.* For some  $t \in [0, 1]$ ,

$$\begin{aligned} |F_\beta(x) - F_\beta(z)| &= |\sum_{j=1}^p \partial_j F_\beta(x + t(z - x))(z_j - x_j)| \\ &\leq \sum_{j=1}^p \pi_j(x + t(z - x)) \max_{1 \leq j \leq p} |z_j - x_j| \leq \max_{1 \leq j \leq p} |z_j - x_j|, \end{aligned}$$

where the property  $\sum_{j=1}^p \pi_j(x + t(z - x)) = 1$  was used. ■

We will also use the following properties of  $m = g \circ F_\beta$ . Here we assume  $g \in C_b^3(\mathbb{R})$  in Lemmas A.4-A.6 below.

**Lemma A.4** (Three derivatives of  $m = g \circ F_\beta$ ). *For every  $1 \leq j, k, l \leq p$ ,*

$$\begin{aligned} \partial_j m(z) &= (\partial g(F_\beta) \pi_j)(z), \\ \partial_j \partial_k m(z) &= (\partial^2 g(F_\beta) \pi_j \pi_k + \partial g(F_\beta) \beta w_{jk})(z), \\ \partial_j \partial_k \partial_l m(z) &= (\partial^3 g(F_\beta) \pi_j \pi_k \pi_l + \partial^2 g(F_\beta) \beta (w_{jk} \pi_l + w_{jl} \pi_k + w_{kl} \pi_j) \\ &\quad + \partial g(F_\beta) \beta^2 q_{jkl})(z), \end{aligned}$$

*where  $\pi_j$ ,  $w_{jk}$  and  $q_{jkl}$  are defined in Lemma A.2, and  $(z)$  denotes evaluation at  $z$ , including evaluation of  $F_\beta$  at  $z$ .*

*Proof of lemma A.4.* The proof follows from repeated application of the chain rule and by the properties noted in Lemma A.2. ■

**Lemma A.5** (Bounds on derivatives of  $m = g \circ F_\beta$ ). *For every  $1 \leq j, k, l \leq p$ ,*

$$|\partial_j \partial_k m(z)| \leq U_{jk}(z), \quad |\partial_j \partial_k \partial_l m(z)| \leq U_{jkl}(z),$$

where

$$\begin{aligned} U_{jk}(z) &:= (G_2 \pi_j \pi_k + G_1 \beta W_{jk})(z), \quad W_{jk}(z) := (\pi_j \delta_{jk} + \pi_j \pi_k)(z), \\ U_{jkl}(z) &:= (G_3 \pi_j \pi_k \pi_l + G_2 \beta (W_{jk} \pi_l + W_{jl} \pi_k + W_{kl} \pi_j) + G_1 \beta^2 Q_{jkl})(z), \\ Q_{jkl}(z) &:= (\pi_j \delta_{jl} \delta_{jk} + \pi_j \pi_l \delta_{jk} + \pi_j \pi_k (\delta_{jl} + \delta_{kl}) + 2\pi_j \pi_k \pi_l)(z). \end{aligned}$$

Moreover,

$$\sum_{j,k=1}^p U_{jk}(z) \leq (G_2 + 2G_1 \beta), \quad \sum_{j,k,l=1}^p U_{jkl}(z) \leq (G_3 + 6G_2 \beta + 6G_1 \beta^2).$$

*Proof of Lemma A.5.* The lemma follows from a direct calculation.  $\blacksquare$

**Lemma A.6** (Stability). *For every  $z \in \mathbb{R}^p$ ,  $w \in \mathbb{R}^p$  such that  $\max_{j \leq p} |w_j| \beta \leq 1$ ,  $\tau \in [0, 1]$ , and every  $1 \leq j, k, l \leq p$ , we have*

$$U_{jk}(z) \lesssim U_{jk}(z + \tau w) \lesssim U_{jk}(z), \quad U_{jkl}(z) \lesssim U_{jkl}(z + \tau w) \lesssim U_{jkl}(z).$$

*Proof of Lemma A.6.* Observe that

$$\pi_j(z + \tau w) = \frac{e^{z_j \beta + \tau w_j \beta}}{\sum_{m=1}^p e^{z_m \beta + \tau w_m \beta}} \leq \frac{e^{z_j \beta}}{\sum_{m=1}^p e^{z_m \beta}} \cdot \frac{e^{\tau \max_{j \leq p} |w_j| \beta}}{e^{-\tau \max_{j \leq p} |w_j| \beta}} \leq e^2 \pi_j(z).$$

Similarly,  $\pi_j(z + \tau w) \geq e^{-2} \pi_j(z)$ . Since  $U_{jk}$  and  $U_{jkl}$  are finite sums of products of terms such as  $\pi_j$ ,  $\pi_k$ ,  $\pi_l$ ,  $\delta_{jk}$ , the claim of the lemma follows.  $\blacksquare$

**A.3. Lemma on Truncation.** The proof of Theorem 2.2 uses the following properties of the truncation operation. Recall that  $\tilde{x}_i = (\tilde{x}_{ij})_{j=1}^p$  and  $\tilde{X} = n^{-1/2} \sum_{i=1}^n \tilde{x}_i$ , where “tilde” denotes the truncation operation defined in Section 2. The following lemma also covers the special case where  $(x_i)_{i=1}^n = (y_i)_{i=1}^n$ . The property (d) is a consequence of sub-Gaussian inequality of [19], Theorem 2.16. for self-normalized sums.

**Lemma A.7** (Truncation Impact). *For every  $1 \leq j, k \leq p$  and  $q \geq 1$ ,*  
 (a)  $(\bar{\mathbb{E}}[|\tilde{x}_{ij}|^q])^{1/q} \leq 2(\bar{\mathbb{E}}[|x_{ij}|^q])^{1/q}$ ; (b)  $\bar{\mathbb{E}}[|\tilde{x}_{ij} \tilde{x}_{ik} - x_{ij} x_{ik}|] \leq (3/2)(\bar{\mathbb{E}}[x_{ij}^2] + \bar{\mathbb{E}}[x_{ik}^2])\varphi(u)$ ; (c)  $\mathbb{E}_n[(\mathbb{E}[x_{ij} - \tilde{x}_{ij}])^2] \leq \bar{\mathbb{E}}[(x_{ij} - \tilde{x}_{ij})^2] \leq \bar{\mathbb{E}}[x_{ij}^2]\varphi^2(u)$ . Moreover, for a given  $\gamma \in (0, 1)$ , let  $u \geq u(\gamma)$  where  $u(\gamma)$  is defined in Section 2. Then:  
 (d) with probability at least  $1 - 5\gamma$ , for all  $1 \leq j \leq p$ ,

$$|X_j - \tilde{X}_j| \leq 5\sqrt{\bar{\mathbb{E}}[x_{ij}^2]}\varphi(u)\sqrt{2\log(p/\gamma)}.$$

*Proof.* See Appendix F.  $\blacksquare$

## APPENDIX B. PROOFS FOR SECTION 2

**B.1. Proof of Theorem 2.1.** Recall that we are assuming that sequences  $(x_i)_{i=1}^n$  and  $(y_i)_{i=1}^n$  are independent. For  $t \in [0, 1]$ , we consider the Slepian interpolation between  $Y$  and  $X$ :

$$Z(t) := \sqrt{t}X + \sqrt{1-t}Y = \sum_{i=1}^n Z_i(t), \quad Z_i(t) := \frac{1}{\sqrt{n}}(\sqrt{t}x_i + \sqrt{1-t}y_i).$$

We shall also employ Stein's leave-one-out expansions:

$$Z^{(i)}(t) := (Z_{ij}(t))_{j=1}^p := Z(t) - Z_i(t).$$

Let  $\Psi(t) = \mathbb{E}[m(Z(t))]$  for  $m := g \circ F_\beta$ . Then by Taylor's theorem,

$$\begin{aligned} \mathbb{E}[m(X) - m(Y)] &= \Psi(1) - \Psi(0) = \int_0^1 \Psi'(t) dt \\ &= \frac{1}{2} \sum_{j=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[\partial_j m(Z(t)) \dot{Z}_{ij}(t)] dt = \frac{1}{2}(I + II + III), \end{aligned}$$

where

$$\begin{aligned} \dot{Z}_{ij}(t) &= \frac{d}{dt} Z_{ij}(t) = \frac{1}{\sqrt{n}} \left( \frac{1}{\sqrt{t}} x_{ij} - \frac{1}{\sqrt{1-t}} y_{ij} \right), \text{ and} \\ I &= \sum_{j=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[\partial_j m(Z^{(i)}(t)) \dot{Z}_{ij}(t)] dt, \\ II &= \sum_{j,k=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[\partial_j \partial_k m(Z^{(i)}(t)) \dot{Z}_{ij}(t) Z_{ik}(t)] dt, \\ III &= \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \int_0^1 (1-\tau) \mathbb{E}[\partial_j \partial_k \partial_l m(Z^{(i)}(t) + \tau Z_i(t)) \dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)] d\tau dt. \end{aligned}$$

Note that random variable  $Z^{(i)}(t)$  and random vector  $(\dot{Z}_{ij}(t), Z_{ij}(t))$  are independent, and  $\mathbb{E}[\dot{Z}_{ij}(t)] = 0$ . Hence we have  $I = 0$ ; moreover, since  $\mathbb{E}[\dot{Z}_{ij}(t) Z_{ik}(t)] = n^{-1} \mathbb{E}[x_{ij} x_{ik} - y_{ij} y_{ik}] = 0$  by construction of  $(y_i)_{i=1}^n$ , we also have  $II = 0$ . Consider the third term  $III$ . We have that

$$\begin{aligned} |III| &\lesssim_{(1)} (G_3 + G_2\beta + G_1\beta^2) n \int \mathbb{E} \left[ \max_{1 \leq j,k,l \leq p} |\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)| \right] dt, \\ &\lesssim_{(2)} n^{-1/2} (G_3 + G_2\beta + G_1\beta^2) \bar{\mathbb{E}} \left[ \max_{1 \leq j \leq p} (|x_{ij}| + |y_{ij}|)^3 \right], \end{aligned}$$

where (1) follows from  $|\partial_j \partial_k \partial_l m(Z^{(i)}(t) + \tau Z_i(t))| \leq U_{jkl}(Z^{(i)}(t) + \tau Z_i(t)) \lesssim (G_3 + G_2\beta + G_1\beta^2)$  holding by Lemma A.5, and (2) is shown below. The first claim of the theorem now follows. The second claim follows directly from property (9) of the smooth max function.

It remains to show (2). Define  $\omega(t) = 1/(\sqrt{t} \wedge \sqrt{1-t})$  and note,

$$\begin{aligned}
& \int_0^1 n \bar{\mathbb{E}} \left[ \max_{1 \leq j, k, l \leq p} |\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)| \right] dt \\
&= \int_0^1 \omega(t) n \bar{\mathbb{E}} \left[ \max_{1 \leq j, k, l \leq p} |\dot{Z}_{ij}(t)/\omega(t) Z_{ik}(t) Z_{il}(t)| \right] dt \\
&\leq n \int_0^1 \omega(t) \left( \bar{\mathbb{E}} \left[ \max_{1 \leq j \leq p} |\dot{Z}_{ij}(t)/\omega(t)|^3 \right] \bar{\mathbb{E}} \left[ \max_{1 \leq j \leq p} |Z_{ij}(t)|^3 \right] \bar{\mathbb{E}} \left[ \max_{1 \leq j \leq p} |Z_{ij}(t)|^3 \right] \right)^{1/3} dt \\
&\leq n^{-1/2} \left\{ \int_0^1 \omega(t) dt \right\} \bar{\mathbb{E}} \left[ \max_{1 \leq j \leq p} (|x_{ij}| + |y_{ij}|)^3 \right]
\end{aligned}$$

where the first inequality follows from Hölder's inequality, and the second from the fact that  $|\dot{Z}_{ij}(t)/\omega(t)| \leq (|x_{ij}| + |y_{ij}|)/\sqrt{n}$ ,  $|Z_{ij}(t)| \leq (|x_{ik}| + |y_{ik}|)/\sqrt{n}$ . Finally we note that  $\int_0^1 \omega(t) dt \lesssim 1$ , so inequality (2) follows. This completes the overall proof.  $\blacksquare$

**B.2. Proof of Corollary 2.1.** In this proof, let  $C > 0$  denote a generic constant depending only on  $c_1$  and  $C_1$ , and its value may change from place to place. For  $\beta > 0$ , define  $e_\beta := \beta^{-1} \log p$ . Recall that  $S_i := \max_{1 \leq j \leq p} (|x_{ij}| + |y_{ij}|)$ . Consider and fix a  $C^3$ -function  $g_0 : \mathbb{R} \rightarrow [0, 1]$  such that  $g_0(s) = 1$  for  $s \leq 0$  and  $g_0(s) = 0$  for  $s \geq 1$ . Fix any  $t \in \mathbb{R}$ , and define  $g(s) = g_0(\psi(s - t - e_\beta))$ . For this function  $g$ ,  $G_0 = 1$ ,  $G_1 \lesssim \psi$ ,  $G_2 \lesssim \psi^2$  and  $G_3 \lesssim \psi^3$ .

Observe now that

$$\begin{aligned}
P(T_0 \leq t) &\leq P(F_\beta(X) \leq t + e_\beta) \leq \mathbb{E}[g(F_\beta(X))] \\
&\leq \mathbb{E}[g(F_\beta(Y))] + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2} \bar{\mathbb{E}}[S_i^3]) \\
&\leq P(F_\beta(Y) \leq t + e_\beta + \psi^{-1}) + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2} \bar{\mathbb{E}}[S_i^3]) \\
&\leq P(Z_0 \leq t + e_\beta + \psi^{-1}) + C(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2} \bar{\mathbb{E}}[S_i^3]).
\end{aligned}$$

where the first inequality follows from (9), the second from construction of  $g$ , the third from Theorem 2.1, and the fourth from construction of  $g$ , and the last from (9). The remaining step is to compare  $P(Z_0 \leq t + e_\beta + \psi^{-1})$  with  $P(Z_0 \leq t)$  and this is where Lemma 2.1 plays its role. By Lemma 2.1,

$$P(Z_0 \leq t + e_\beta + \psi^{-1}) - P(Z_0 \leq t) \leq C(e_\beta + \psi^{-1}) \sqrt{1 \vee \log(p\psi)}.$$

by which we have

$$P(T_0 \leq t) - P(Z_0 \leq t) \leq C[(\psi^3 + \beta\psi^2 + \beta^2\psi)(n^{-1/2} \bar{\mathbb{E}}[S_i^3]) + (e_\beta + \psi^{-1}) \sqrt{1 \vee \log(p\psi)}].$$

We have to minimize the right side with respect to  $\beta$  and  $\psi$ . It is reasonable to choose  $\beta$  in such a way that  $e_\beta$  and  $\psi^{-1}$  are balanced, i.e.,  $\beta = \psi \log p$ . With this  $\beta$ , the bracket on the right side is

$$\lesssim \psi^3 (\log p)^2 (n^{-1/2} \bar{\mathbb{E}}[S_i^3]) + \psi^{-1} \sqrt{1 \vee \log(p\psi)},$$

which is approximately minimized by  $\psi = (\log p)^{-3/8} (n^{-1/2} \bar{\mathbb{E}}[S_i^3])^{-1/4}$ . With this  $\psi$ ,  $\psi \leq (n^{-1/2} \bar{\mathbb{E}}[S_i^3])^{-1/4} \leq Cn^{1/8}$  (recall that  $p \geq 3$ ), and hence  $\log(p\psi) \leq C \log(pn)$ . Therefore,

$$P(T_0 \leq t) - P(Z_0 \leq t) \leq C(n^{-1/2} \bar{\mathbb{E}}[S_i^3])^{1/4} (\log(pn))^{7/8}.$$

This gives one half of the claim. The other half follows similarly.  $\blacksquare$

**B.3. Proof of Theorem 2.2.** The second claim of the theorem follows from property (9) of the smooth max function. Hence we shall prove the first claim. The proof strategy is similar to the proof of Theorem 2.1. However, to control effectively the third order terms in the leave-one-out expansions we shall use truncation and replace  $X$  and  $Y$  by their truncated versions  $\tilde{X}$  and  $\tilde{Y}$ , defined as follows: let  $\tilde{x}_i = (\tilde{x}_{ij})_{j=1}^p$ , where  $\tilde{x}_{ij}$  was defined before the statement of the theorem, and define the truncated version of  $X$  as  $\tilde{X} = n^{-1/2} \sum_{i=1}^n \tilde{x}_i$ . Also let

$$\tilde{y}_i := (\tilde{y}_{ij})_{j=1}^p, \quad \tilde{y}_{ij} := y_{ij} 1 \left\{ |y_{ij}| \leq u(\bar{\mathbb{E}}[y_{ij}^2])^{1/2} \right\}, \quad \tilde{Y} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{y}_i.$$

Note that by the symmetry of the distribution of  $y_{ij}$ ,  $\mathbb{E}[\tilde{y}_{ij}] = 0$ . Recall that we are assuming that sequences  $(x_i)_{i=1}^n$  and  $(y_i)_{i=1}^n$  are independent.

The proof consists of four steps. Step 1 will show that we can replace  $X$  by  $\tilde{X}$  and  $Y$  by  $\tilde{Y}$ . Step 2 will bound the difference of the expectations of the relevant functions of  $\tilde{X}$  and  $\tilde{Y}$ . This is the main step of the proof. Steps 3 and 4 will carry out supporting calculations. The steps of the proof will also call on various technical lemmas collected in Appendix A.

**Step 1.** Let  $m := g \circ F_\beta$ . The main goal is to bound  $\mathbb{E}[m(X) - m(Y)]$ . Define

$$\mathcal{I} = 1 \left\{ \max_{1 \leq j \leq p} |X_j - \tilde{X}_j| \leq \Delta(\gamma, u) \text{ and } \max_{1 \leq j \leq p} |Y_j - \tilde{Y}_j| \leq \Delta(\gamma, u) \right\},$$

where  $\Delta(\gamma, u) := 5M_2\varphi(u)\sqrt{2\log(p/\gamma)}$ . By Lemma A.7 we have  $\mathbb{E}[\mathcal{I}] \geq 1 - 10\gamma$ . Observe that by Lemma A.3,

$$|m(x) - m(y)| \leq G_1 |F_\beta(x) - F_\beta(y)| \leq G_1 \max_{1 \leq j \leq p} |x_j - y_j|,$$

so that

$$\begin{aligned} |\mathbb{E}[m(X) - m(\tilde{X})]| &\leq |\mathbb{E}[(m(X) - m(\tilde{X}))\mathcal{I}]| + |\mathbb{E}[(m(X) - m(\tilde{X}))(1 - \mathcal{I})]| \\ &\lesssim G_1 \Delta(\gamma, u) + G_0 \gamma, \\ |\mathbb{E}[m(Y) - m(\tilde{Y})]| &\leq |\mathbb{E}[(m(Y) - m(\tilde{Y}))\mathcal{I}]| + |\mathbb{E}[(m(Y) - m(\tilde{Y}))(1 - \mathcal{I})]| \\ &\lesssim G_1 \Delta(\gamma, u) + G_0 \gamma, \end{aligned}$$

and hence

$$|\mathbb{E}[m(X) - m(Y)]| \lesssim |\mathbb{E}[m(\tilde{X}) - m(\tilde{Y})]| + G_1 \Delta(\gamma, u) + G_0 \gamma.$$

**Step 2.** (Main Step) The purpose of this step is to establish the bound:

$$|\mathbb{E}[m(\tilde{X}) - m(\tilde{Y})]| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)M_3^3 + (G_2 + \beta G_1)M_2^2\varphi(u).$$

Define, as in the proof of Theorem 2.1,

$$Z(t) := \sqrt{t}\tilde{X} + \sqrt{1-t}\tilde{Y} = \sum_{i=1}^n Z_i(t), \quad Z_i(t) := \frac{1}{\sqrt{n}}(\sqrt{t}\tilde{x}_i + \sqrt{1-t}\tilde{y}_i), \text{ and}$$

$$Z^{(i)}(t) := Z(t) - Z_i(t), \quad \dot{Z}_{ij}(t) = \frac{1}{\sqrt{n}} \left( \frac{1}{\sqrt{t}}\tilde{x}_{ij} - \frac{1}{\sqrt{1-t}}\tilde{y}_{ij} \right).$$

Arguing as in the proof of Theorem 2.1, we have

$$\mathbb{E}[m(\tilde{X}) - m(\tilde{Y})] = \frac{1}{2} \sum_{j=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[\partial_j m(Z(t)) \dot{Z}_{ij}(t)] dt = \frac{1}{2}(I + II + III),$$

where

$$I = \sum_{j=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[\partial_j m(Z^{(i)}(t)) \dot{Z}_{ij}(t)] dt,$$

$$II = \sum_{j,k=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[\partial_j \partial_k m(Z^{(i)}(t)) \dot{Z}_{ij}(t) Z_{ik}(t)] dt,$$

$$III = \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \int_0^1 (1-\tau) \mathbb{E}[\partial_j \partial_k \partial_l m(Z^{(i)}(t) + \tau Z_i(t)) \dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)] d\tau dt.$$

By independence of  $Z^{(i)}(t)$  and  $\dot{Z}_{ij}(t)$  together with the fact that  $\mathbb{E}[\dot{Z}_{ij}(t)] = 0$ , we have  $I = 0$ . Moreover, in steps 3 and 4 below, we will show that

$$|II| \lesssim (G_2 + \beta G_1)M_2^2\varphi(u), \quad |III| \lesssim n^{-1/2}(G_3 + G_2\beta + G_1\beta^2)M_3^3.$$

The claim of this step now follows.

**Step 3.** (Bound on  $II$ ) By independence of  $Z^{(i)}(t)$  and  $\dot{Z}_{ij}(t)Z_{ik}(t)$ ,

$$\begin{aligned} |II| &= \left| \sum_{j,k=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[\partial_j \partial_k m(Z^{(i)}(t))] \mathbb{E}[\dot{Z}_{ij}(t) Z_{ik}(t)] dt \right| \\ &\leq \sum_{j,k=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[|\partial_j \partial_k m(Z^{(i)}(t))|] \cdot |\mathbb{E}[\dot{Z}_{ij}(t) Z_{ik}(t)]| dt \\ &\leq \sum_{j,k=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[U_{jk}(Z^{(i)}(t))] \cdot |\mathbb{E}[\dot{Z}_{ij}(t) Z_{ik}(t)]| dt, \end{aligned}$$

where the last step follows from Lemma A.5. Since  $|\sqrt{t}\tilde{x}_{ij} + \sqrt{1-t}\tilde{y}_{ij}| \leq 2\sqrt{2}uM_2$ , so that  $|\beta(\sqrt{t}\tilde{x}_{ij} + \sqrt{1-t}\tilde{y}_{ij})/\sqrt{n}| \leq 1$  (which is satisfied by the assumption  $\beta 2\sqrt{2}uM_2/\sqrt{n} \leq 1$ ), by Lemmas A.6 and A.5, the last expression



is bounded by

$$\begin{aligned}
& \sum_{j,k=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[U_{jk}(Z(t))] \cdot |\mathbb{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]| dt \\
&= \int_0^1 \left\{ \sum_{j,k=1}^p \mathbb{E}[U_{jk}(Z(t))] \right\} \sum_{i=1}^n |\mathbb{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]| dt \\
&\lesssim (G_2 + G_1\beta) \int_0^1 \sum_{i=1}^n |\mathbb{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]| dt.
\end{aligned}$$

Observe that since  $\mathbb{E}[x_{ij}x_{ik}] = \mathbb{E}[y_{ij}y_{ik}]$ , we have that  $\mathbb{E}[\dot{Z}_{ij}(t)Z_{ik}(t)] = n^{-1}\mathbb{E}[\tilde{x}_{ij}\tilde{x}_{ik} - \tilde{y}_{ij}\tilde{y}_{ik}] = n^{-1}\mathbb{E}[\tilde{x}_{ij}\tilde{x}_{ik} - x_{ij}x_{ik}] + n^{-1}\mathbb{E}[y_{ij}y_{ik} - \tilde{y}_{ij}\tilde{y}_{ik}]$ , so that by Lemma A.7 (b),  $\sum_{i=1}^n |\mathbb{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]| \leq \bar{\mathbb{E}}[|\tilde{x}_{ij}\tilde{x}_{ik} - x_{ij}x_{ik}|] + \bar{\mathbb{E}}[|y_{ij}y_{ik} - \tilde{y}_{ij}\tilde{y}_{ik}|] \lesssim (\bar{\mathbb{E}}[x_{ij}^2] + \bar{\mathbb{E}}[x_{ik}^2])\varphi(u) \lesssim M_2^2\varphi(u)$ . Therefore, we conclude that  $|II| \lesssim (G_2 + G_1\beta)M_2^2\varphi(u)$ .

**Step 4.** (Bound on  $III$ ) Observe that

$$\begin{aligned}
|III| &\leq_{(1)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \int_0^1 \mathbb{E}[U_{jkl}(Z^{(i)}(t) + \tau Z_i(t)) |\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|] d\tau dt \\
&\lesssim_{(2)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[U_{jkl}(Z^{(i)}(t)) |\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|] dt \\
(20) \quad &=_{(3)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[U_{jkl}(Z^{(i)}(t))] \cdot \mathbb{E}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|] dt,
\end{aligned}$$

where (1) follows from  $|\partial_j \partial_k \partial_l m(z)| \leq U_{jkl}(z)$  (see Lemma A.5), (2) from Lemma A.6, (3) from independence of  $Z^{(i)}(t)$  and  $\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)$ . Moreover, the last expression is bounded as follows:

$$\begin{aligned}
\text{right side of (20)} &\lesssim_{(4)} \sum_{j,k,l=1}^p \sum_{i=1}^n \int_0^1 \mathbb{E}[U_{jkl}(Z(t))] \cdot \mathbb{E}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|] dt \\
&=_{(5)} \sum_{j,k,l=1}^p \int_0^1 \mathbb{E}[U_{jkl}(Z(t))] \cdot n\bar{\mathbb{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|] dt \\
&\leq_{(6)} \int_0^1 \left( \sum_{j,k,l=1}^p \mathbb{E}[U_{jkl}(Z(t))] \right) \max_{1 \leq j,k,l \leq p} n\bar{\mathbb{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|] dt \\
&\lesssim_{(7)} (G_3 + G_2\beta + G_1\beta^2) \int_0^1 \max_{1 \leq j,k,l \leq p} n\bar{\mathbb{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|] dt,
\end{aligned}$$

where (4) follows from Lemma A.6, (5) from definition of  $\bar{E}$ , (6) from a trivial inequality, (7) from Lemma A.5. We have to bound the integral on the last line. Let  $\omega(t) = 1/(\sqrt{t} \wedge \sqrt{1-t})$ , and observe that

$$\begin{aligned} & \int_0^1 \max_{1 \leq j, k, l \leq p} n \bar{E}[|\dot{Z}_{ij}(t) Z_{ik}(t) Z_{il}(t)|] dt \\ &= \int_0^1 \omega(t) \max_{1 \leq j, k, l \leq p} n \bar{E}[|(\dot{Z}_{ij}(t)/\omega(t)) Z_{ik}(t) Z_{il}(t)|] dt \\ &\leq n \int_0^1 \omega(t) \max_{1 \leq j, k, l \leq p} \left( \bar{E}[|\dot{Z}_{ij}(t)/\omega(t)|^3] \bar{E}[|Z_{ik}(t)|^3] \bar{E}[|Z_{il}(t)|^3] \right)^{1/3} dt, \end{aligned}$$

where the last inequality is by Hölder. The last term is further bounded as

$$\begin{aligned} &\leq_{(1)} n^{-1/2} \left\{ \int_0^1 \omega(t) dt \right\} \max_{1 \leq j \leq p} \bar{E}[ (|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)^3 ] \\ &\lesssim_{(2)} n^{-1/2} \max_{1 \leq j \leq p} [(\bar{E}[|\tilde{x}_{ij}|^3])^{1/3} + (\bar{E}[|\tilde{y}_{ij}|^3])^{1/3}]^3 \\ &\lesssim_{(3)} n^{-1/2} \max_{1 \leq j \leq p} [(\bar{E}[|x_{ij}|^3])^{1/3} + (\bar{E}[|y_{ij}|^3])^{1/3}]^3 \\ &\lesssim_{(4)} n^{-1/2} \max_{1 \leq j \leq p} \bar{E}[|x_{ij}|^3], \end{aligned}$$

where (1) follows from the fact that:  $|\dot{Z}_{ij}(t)/\omega(t)| \leq (|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)/\sqrt{n}$ ,  $|Z_{im}(t)| \leq (|\tilde{x}_{im}| + |\tilde{y}_{im}|)/\sqrt{n}$ , and the product of terms  $\bar{E}[ (|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)^3 ]^{1/3}$ ,  $\bar{E}[ (|\tilde{x}_{ik}| + |\tilde{y}_{ik}|)^3 ]^{1/3}$  and  $\bar{E}[ (|\tilde{x}_{il}| + |\tilde{y}_{il}|)^3 ]^{1/3}$  is trivially bounded by  $\max_{1 \leq j \leq p} \bar{E}[ (|\tilde{x}_{ij}| + |\tilde{y}_{ij}|)^3 ]$ ; (2) follows from  $\int_0^1 \omega(t) dt \lesssim 1$ , (3) from Lemma A.7 (a), and (4) from the normality of  $y_{ij}$  with  $E[y_{ij}^2] = E[x_{ij}^2]$ , so that  $E[|y_{ij}|^3] \lesssim (E[y_{ij}^2])^{3/2} = (E[x_{ij}^2])^{3/2} \leq E[|x_{ij}|^3]$ . This completes the overall proof. ■

**B.4. Proof of Corollary 2.2.** See Supplemental Appendix F.2. ■

**B.5. Proof of Lemma 2.2.** Since  $\bar{E}[x_{ij}^2] \geq c_1$  by assumption, we have  $1\{|x_{ij}| > u(\bar{E}[x_{ij}^2])^{1/2}\} \leq 1\{|x_{ij}| > c_1^{1/2}u\}$ . By Markov's inequality and the condition of the lemma, we have

$$\begin{aligned} &P\left(|x_{ij}| > u(\bar{E}[x_{ij}^2])^{1/2}, \text{ for some } (i, j)\right) \leq \sum_{i=1}^n P\left(\max_{1 \leq j \leq p} |x_{ij}| > c_1^{1/2}u\right) \\ &\leq \sum_{i=1}^n P\left(h(\max_{1 \leq j \leq p} |x_{ij}|/D) > h(c_1^{1/2}u/D)\right) \leq n/h(c_1^{1/2}u/D). \end{aligned}$$

This implies  $u_x(\gamma) \leq c_1^{-1/2} D h^{-1}(n/\gamma)$ . For  $u_y(\gamma)$ , by  $y_{ij} \sim N(0, E[x_{ij}^2])$  with  $E[x_{ij}^2] \leq B^2$ , we have  $E[\exp(y_{ij}^2/(4B^2))] \lesssim 1$ . Hence

$$\begin{aligned} &P\left(|y_{ij}| > u(\bar{E}[y_{ij}^2])^{1/2}, \text{ for some } (i, j)\right) \leq \sum_{i=1}^n \sum_{j=1}^p P(|y_{ij}| > c_1^{1/2}u) \\ &\leq \sum_{i=1}^n \sum_{j=1}^p P(|y_{ij}|/(2B) > c_1^{1/2}u/(2B)) \lesssim np \exp(-c_1 u^2/(4B^2)). \end{aligned}$$

Therefore,  $u_y(\gamma) \leq CB\sqrt{\log(pn/\gamma)}$  where  $C > 0$  depends only on  $c_1$ . ■

**B.6. Proof of Corollary 2.3.** Case (i) follows directly from Corollary 2.1. Hence we only consider cases (ii)-(v).

**Step 1.** In this step, in each case of conditions (E.2)-(E.5), we shall compute the following bounds on moments  $M_3$  and  $M_4$  and parameters  $B$  and  $D$  in Lemma 2.2 with specific choice of  $h$ :

$$\begin{aligned} \text{(E.2)} \quad & B \vee M_3^3 \vee M_4^2 \leq C, \quad D \leq C \log p, \quad h(v) = e^v - 1; \\ \text{(E.3)} \quad & B = B_n, \quad D \leq CB_n, \quad M_3^3 \vee M_4^2 \leq CB_n, \quad h(v) = e^v - 1; \\ \text{(E.4)} \quad & B \vee M_3^3 \vee M_4^2 \leq CB_n, \quad D \leq CB_n \log p, \quad h(v) = e^v - 1; \\ \text{(E.5)} \quad & B \vee D \vee M_3^3 \vee M_4^2 \leq CB_n, \quad h(v) = v^4. \end{aligned}$$

Here  $C > 0$  is a (sufficiently large) constant that depends only on  $c_1$  and  $C_1$ . The bounds on  $B$ ,  $M_3$  and  $M_4$  follow from elementary computations using Hölder's inequality. The bounds on  $D$  follow from an elementary application of Lemma 2.2.2 in [44]. For brevity, we omit the detail.

**Step 2.** In either case of (ii)-(v), there are sufficiently small constants  $c_3 > 0$  and  $c_4 > 0$ , and a sufficiently large constant  $C_2 > 0$ , depending only on  $c_1, c_2, C_1$  such that, with  $\ell_n := \log(pn^{1+c_3})$ ,

$$\begin{aligned} n^{-1/2} \ell_n^{3/2} \max\{B \ell_n^{1/2}, D h^{-1}(n^{1+c_3})\} &\leq C_2 n^{-c_4}, \\ n^{-1/8} (M_3^{3/4} \vee M_4^{1/2}) \ell_n^{7/8} &\leq C_2 n^{-c_4}. \end{aligned}$$

Hence taking  $\gamma = n^{-c_3}$ , we conclude from Corollary 2.2 and Lemma 2.2 that  $\rho \leq C n^{-\min\{c_3, c_4\}}$  where  $C > 0$  depends only on  $c_1, c_2, C_1$ .  $\blacksquare$

## APPENDIX C. PROOFS FOR SECTION 3

**C.1. Proof of Lemma 3.2.** Recall that  $\Delta = \max_{1 \leq j, k \leq p} |\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathbb{E}}[x_{ij}x_{ik}]|$ . By Lemma 3.1, on the event  $\{(x_i)_{i=1}^n : \Delta \leq \vartheta\}$ , we have  $|\mathbb{P}(Z_0 \leq t) - \mathbb{P}_e(W_0 \leq t)| \leq \pi(\vartheta)$  for all  $t \in \mathbb{R}$ , and so on this event

$$\mathbb{P}_e(W_0 \leq c_{Z_0}(\alpha + \pi(\vartheta))) \geq \mathbb{P}(Z_0 \leq c_{Z_0}(\alpha + \pi(\vartheta))) - \pi(\vartheta) \geq \alpha + \pi(\vartheta) - \pi(\vartheta) = \alpha,$$

implying the first claim. The second claim follows similarly.  $\blacksquare$

**C.2. Proof of Lemma 3.3.** By equation (15), the probability of the event  $\{(x_i)_{i=1}^n : \mathbb{P}_e(|W - W_0| > \zeta_1) \leq \zeta_2\}$  is at least  $1 - \zeta_2$ . On this event,

$$\mathbb{P}_e(W \leq c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geq \mathbb{P}_e(W_0 \leq c_{W_0}(\alpha + \zeta_2)) - \zeta_2 \geq \alpha + \zeta_2 - \zeta_2 = \alpha,$$

implying that  $\mathbb{P}(c_W(\alpha) \leq c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geq 1 - \zeta_2$ . The second claim of the lemma follows similarly.  $\blacksquare$

**C.3. Proof of Theorem 3.1.** For  $\vartheta > 0$ , let  $\pi(\vartheta) := C_2 \vartheta^{1/3} (1 \vee \log(p/\vartheta))^{2/3}$  as defined in Lemma 3.2. Then

$$\begin{aligned} \mathbb{P}(T_0 \leq c_{W_0}(\alpha)) &\leq_{(1)} \mathbb{P}(T_0 \leq c_{Z_0}(\alpha + \pi(\vartheta))) + \mathbb{P}(\Delta > \vartheta) \\ &\leq_{(2)} \alpha + \pi(\vartheta) + \mathbb{P}(\Delta > \vartheta) + \rho, \end{aligned}$$

where (1) follows from Lemma 3.2 and (2) follows from definition of  $\rho$  and the fact that  $Z_0$  has no point masses. The upper bound is proven. The lower bound follows similarly.  $\blacksquare$

**C.4. Proof of Theorem 3.2.** For  $\vartheta > 0$ , let  $\pi(\vartheta) := C_2 \vartheta^{1/3} (1 \vee \log(p/\vartheta))^{2/3}$  with  $C_2 > 0$  as in Lemma 3.2. Then

$$\begin{aligned}
& \mathbb{P}(T \leq c_W(\alpha)) \leq_{(1)} \mathbb{P}(T_0 \leq c_W(\alpha) + \zeta_1) + \zeta_2 \\
& \leq_{(2)} \mathbb{P}(T_0 \leq c_{W_0}(\alpha + \zeta_2) + 2\zeta_1) + 2\zeta_2 \\
& \leq_{(3)} \mathbb{P}(T_0 \leq c_{Z_0}(\alpha + \zeta_2 + \pi(\vartheta)) + 2\zeta_1) + 2\zeta_2 + \mathbb{P}(\Delta > \vartheta) \\
& \leq_{(4)} \mathbb{P}(Z_0 \leq c_{Z_0}(\alpha + \zeta_2 + \pi(\vartheta)) + 2\zeta_1) + \rho + 2\zeta_2 + \mathbb{P}(\Delta > \vartheta) \\
& \leq_{(5)} \mathbb{P}(Z_0 \leq c_{Z_0}(\alpha + \zeta_2 + \pi(\vartheta))) + C_3 \zeta_1 \sqrt{1 \vee \log(p/\zeta_1)} + \rho + 2\zeta_2 + \mathbb{P}(\Delta > \vartheta) \\
& \leq_{(6)} \alpha + \zeta_2 + \pi(\vartheta) + C_3 \zeta_1 \sqrt{1 \vee \log(p/\zeta_1)} + 2\zeta_2 + \mathbb{P}(\Delta > \vartheta) + \rho
\end{aligned}$$

where  $C_3 > 0$  depends on  $c_1$  and  $C_1$  only and where (1) follows from equation (14), (2) from Lemma 3.3, (3) from Lemma 3.2, (4) from the definition of  $\rho$ , and (5) follows from Lemma 2.1 on anti-concentration, and (6) by the fact that  $Z_0$  has no point masses. This gives the upper bound. The lower bound follows similarly.  $\blacksquare$

**C.5. Proof of Corollary 3.1.** The proof of this corollary relies on:

**Lemma C.1.** *Recall conditions (E.2)-(E.5) in Section 2.2. Then*

$$\mathbb{E}[\Delta] \leq C \times \begin{cases} \sqrt{\frac{\log p}{n}} \sqrt{\frac{(\log(pn))^2 (\log p)}{n}}, & \text{under (E.2),} \\ \sqrt{\frac{B_n^2 \log p}{n}} \sqrt{\frac{B_n^2 (\log p)}{n}}, & \text{under (E.3),} \\ \sqrt{\frac{B_n^2 \log p}{n}} \sqrt{\frac{B_n^2 (\log(pn))^2 (\log p)}{n}}, & \text{under (E.4),} \\ \sqrt{\frac{B_n^2 \log p}{n}} \sqrt{\frac{B_n^2 (\log p)}{\sqrt{n}}}, & \text{under (E.5),} \end{cases}$$

where  $C > 0$  depends only on  $c_1$  and  $C_1$  that appear in (E.2)-(E.5).

*Proof.* By Lemma A.1 and Hölder's inequality, we have

$$\mathbb{E}[\Delta] \lesssim M_4^2 \sqrt{(\log p)/n} + (\mathbb{E}[\max_{i,j} |x_{ij}|^4])^{1/2} (\log p)/n.$$

The conclusion of the lemma follows from elementary calculations with help of Lemma 2.2.2 in [44].  $\blacksquare$

*Proof of Corollary 3.1.* We make use of Theorem 3.2. Let  $c > 0$  and  $C > 0$  denote generic constants depending only on  $c_1, c_2, C_1$ , and their values may change from place to place. By Corollary 2.3, in either case of (i)-(iv),  $\rho \leq Cn^{-c}$ . Moreover,  $\zeta_1 \sqrt{\log p} \leq C_1 n^{-c_2}$  implies that  $\zeta_1 \leq C_1 n^{-c_2}$  (recall  $p \geq 3$ ), and hence  $\zeta_1 \sqrt{\log(p/\zeta_1)} \leq Cn^{-c}$ . Also,  $\zeta_2 \leq Cn^{-c}$  by assumption.

Let  $\vartheta = \vartheta_n := (\mathbb{E}[\Delta])^{1/2} / \log p$ . By Lemma C.1,  $\mathbb{E}[\Delta] (\log p)^2 \leq Cn^{-c}$ . Therefore,  $\pi(\vartheta) \leq Cn^{-c}$  (with possibly different  $c, C > 0$ ). In addition, by Markov's inequality,  $\mathbb{P}(\Delta > \vartheta) \leq \mathbb{E}[\Delta] / \vartheta \leq Cn^{-c}$ . Hence, by Theorem 3.2, we have  $\sup_{\alpha \in (0,1)} |\mathbb{P}(T \leq c_W(\alpha)) - \alpha| \leq Cn^{-c}$ .  $\blacksquare$

## APPENDIX D. PROOFS FOR SECTION 4

**D.1. Proof of Theorem 4.1.** The proof proceeds in three steps. In the proof  $(\hat{\beta}, \lambda)$  denotes  $(\hat{\beta}^{(k)}, \lambda^{(k)})$  with  $k$  either 0 or 1.

**Step 1.** Here we show that there exist some constants  $c > 0$  and  $C > 0$  (depending only  $c_1, C_1$  and  $\sigma^2$ ) such that for either  $k \in \{0, 1\}$ ,

$$(21) \quad P(T_0 \leq \lambda^{(k)}) \geq 1 - \alpha - \nu_n,$$

with  $\nu_n = Cn^{-c}$ . We first note that  $T_0 = \sqrt{n} \max_{1 \leq k \leq 2p} \mathbb{E}_n[\tilde{z}_{ik} \varepsilon_i]$ , where  $\tilde{z}_i = (z'_i, -z'_i)'$ . Application of Corollary 2.3-(v) gives

$$|P(T_0 \leq \lambda) - P(Z_0 \leq \lambda)| \leq Cn^{-c},$$

where  $c > 0$  and  $C > 0$  are constants depending only on  $c_1, C_1$  and  $\sigma^2$ . Since  $\lambda \geq c_{Z_0}(1 - \alpha)$ , the claim follows. Indeed,  $\lambda^{(1)} = c_{Z_0}(1 - \alpha)$ , and  $\lambda^{(1)} \leq \lambda^{(0)} = c_0(1 - \alpha) := \sigma \Phi^{-1}(1 - \alpha/(2p))$ , since by the union bound  $P(Z_0 \geq c_0(1 - \alpha)) \leq 2pP(\sigma N(0, 1) \geq c_0(1 - \alpha)) = \alpha$ .

**Step 2.** We claim that with probability  $\geq 1 - \alpha - \nu_n$ ,  $\hat{\delta} = \hat{\beta} - \beta$  obeys:

$$\sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}(z'_i \hat{\delta})]| \leq 2\lambda.$$

Indeed, by definition of  $\hat{\beta}$ ,  $\sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}(y_i - z'_i \hat{\beta})]| \leq \lambda$ , which by the triangle inequality implies  $\sqrt{n} \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}(z'_i \hat{\delta})]| \leq T_0 + \lambda$ . The claim follows from Step 1.

**Step 3.** By Step 1, with probability  $\geq 1 - \alpha - \nu_n$ , the true value  $\beta$  obeys the constraint in optimization problem (16), in which case by definition of  $\hat{\beta}$ ,  $\|\hat{\beta}\|_{\ell_1} \leq \|\beta\|_{\ell_1}$ . Therefore, with the same probability,  $\hat{\delta} \in \mathcal{R}(\beta) = \{\delta \in \mathbb{R}^d : \|\beta + \delta\|_{\ell_1} \leq \|\beta\|_{\ell_1}\}$ . By definition of  $\kappa_I(\beta)$  we have that

$$\kappa_I(\beta) \|\hat{\delta}\|_I \leq \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}(z'_i \hat{\delta})]|.$$

Combining this inequality with Step 2 gives the claim of the theorem.  $\blacksquare$

**D.2. Proof of Theorem 4.2.** The proof has four steps. In the proof, we let  $\varrho_n = Cn^{-c}$  for sufficiently small  $c > 0$  and sufficiently large  $C > 0$  depending only on  $c_1, C_1, \underline{\sigma}^2, \sigma^2$ , where  $c$  and  $C$  (and hence  $\varrho_n$ ) may change from place to place.

**Step 0.** The same argument as in the previous proof applies to  $\hat{\beta}^{(0)}$  with  $\lambda = \lambda^{(0)} := c_0(1 - 1/n)$ , where now  $\sigma^2$  is the upper bound on  $E[\varepsilon_i^2]$ . Thus, we conclude that with probability at least  $1 - \varrho_n$ ,

$$\|\hat{\beta}^{(0)} - \beta\|_{\text{pr}} \leq \frac{2c_0(1 - 1/n)}{\sqrt{n}\kappa_{\text{pr}}(\beta)}.$$

**Step 1.** We claim that with probability at least  $1 - \varrho_n$ ,

$$\max_{1 \leq j \leq p} (\mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i - \varepsilon_i)^2])^{1/2} \leq B_n \frac{2c_0(1 - 1/n)}{\sqrt{n}\kappa_{\text{pr}}(\beta)} =: \iota_n.$$

Application of Hölder's inequality and identity  $\varepsilon_i - \widehat{\varepsilon}_i = z'_i(\widehat{\beta}^{(0)} - \beta)$  gives

$$\max_{1 \leq j \leq p} (\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2])^{1/2} \leq B_n(\mathbb{E}_n[z'_i(\widehat{\beta}^{(0)} - \beta)]^2)^{1/2} \leq B_n\|\widehat{\beta}^{(0)} - \beta\|_{\text{pr}}.$$

The claim follows from Step 0.

**Step 2.** In this step, we apply Corollary 3.1-(iv) to

$$\begin{aligned} T &= T_0 = \sqrt{n} \max_{1 \leq j \leq 2p} \mathbb{E}_n[\tilde{z}_{ij}\varepsilon_i], \quad W = \sqrt{n} \max_{1 \leq j \leq 2p} \mathbb{E}_n[\tilde{z}_{ij}\widehat{\varepsilon}_i e_i], \quad \text{and} \\ W_0 &= \sqrt{n} \max_{1 \leq j \leq 2p} \mathbb{E}_n[\tilde{z}_{ij}\varepsilon_i e_i], \end{aligned}$$

where  $\tilde{z}_i = (z'_i, -z'_i)'$ , to conclude that uniformly in  $\alpha \in (0, 1)$

$$(22) \quad \mathbb{P}(T_0 \leq c_W(1 - \alpha)) \geq 1 - \alpha - \varrho_n.$$

To show applicability of Corollary 3.1-(iv), we note that for any  $\zeta_1 > 0$ ,

$$\begin{aligned} \mathbb{P}_e(|W - W_0| > \zeta_1) &\leq \mathbb{E}_e[\|W - W_0\|/\zeta_1] \leq \sqrt{n}\mathbb{E}_e \left[ \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}(\widehat{\varepsilon}_i - \varepsilon_i)e_i]| \right] / \zeta_1 \\ &\lesssim \sqrt{\log p} \max_{1 \leq j \leq p} (\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2])^{1/2} / \zeta_1, \end{aligned}$$

where the third inequality is due to Pisier's inequality. The last quantity is bounded by  $(\iota_n^2 \log p)^{1/2} / \zeta_1$  with probability  $\geq 1 - \varrho_n$  by Step 1.

Since  $\iota_n \log p \leq C_1 n^{-c_1}$  by assumption (vi) of the theorem, we can take  $\zeta_1$  in such a way that  $\zeta_1(\log p)^{1/2} \leq \varrho_n$  and  $(\iota_n^2 \log p)^{1/2} / \zeta_1 \leq \varrho_n$ . Then all the conditions of Corollary 3.1-(iv) with so defined  $\zeta_1$  and  $\zeta_2 = \varrho_n \vee ((\iota_n^2 \log p)^{1/2} / \zeta_1)$  are satisfied, and hence application of the corollary gives that uniformly in  $\alpha \in (0, 1)$ ,

$$(23) \quad |\mathbb{P}(T_0 \leq c_W(1 - \alpha)) - 1 - \alpha| \leq \varrho_n,$$

which implies the claim of this step.

**Step 3.** In this step we claim that with probability at least  $1 - \varrho_n$ ,

$$c_W(1 - \alpha) \leq c_{Z_0}(1 - \alpha + 2\varrho_n).$$

Combining Step 2 and Lemma 3.3 gives that with probability at least  $1 - \zeta_2$ ,  $c_W(1 - \alpha) \leq c_{W_0}(1 - \alpha + \zeta_2) + \zeta_1$ , where  $\zeta_1$  and  $\zeta_2$  are chosen as in Step 2. In addition, Lemma 3.2 shows that  $c_{W_0}(1 - \alpha + \zeta_2) \leq c_{Z_0}(1 - \alpha + \varrho_n)$ . Finally, Lemma 2.1 yields  $c_{Z_0}(1 - \alpha + \varrho_n) + \zeta_1 \leq c_{Z_0}(1 - \alpha + 2\varrho_n)$ . Combining these bounds gives the claim of this step.

**Step 4.** Given (22), the rest of the proof is identical to Steps 2-3 in the proof of Theorem 4.1 with  $\lambda = c_W(1 - \alpha)$ . The result follows for  $\nu_n = 2\varrho_n$ . ■

#### ACKNOWLEDGMENTS

The authors would like to express their appreciation to L.H.Y. Chen, David Gamarnik, Qi-Man Shao, Vladimir Koltchinskii, Enno Mammen, Axel Munk, and Steve Portnoy for enlightening discussions.

## REFERENCES

- [1] Alquier, P. and Hebiri, M. (2011). Generalization of  $\ell_1$  constraints for high dimensional regression problems. *Statist. Probab. Lett.* **81** 1760-1765.
- [2] Arlot, S., Blanchard, G. and Roquain, E. (2010a). Some non-asymptotic results on resampling in high dimension I: confidence regions. *Ann. Statist.* **38** 51-82.
- [3] Arlot, S., Blanchard, G. and Roquain, E. (2010b). Some non-asymptotic results on resampling in high dimension II: multiple tests. *Ann. Statist.* **38** 83-99.
- [4] Ball, K. (1993). The reverse isoperimetric problem for Gaussian measure. *Discrete Comput. Geom.* **10** 411-420.
- [5] Belloni, A. and Chernozhukov, V. (2009). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, to appear.
- [6] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791-806.
- [7] Bentkus, V. (2003). On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Infer.* **113** 385-402.
- [8] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705-1732.
- [9] Bretagnolle, J. and Massart, P. (1989). Hungarian construction from the non asymptotic viewpoint. *Ann. Probab.* **17** 239-256.
- [10] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.
- [11] Candès, E.J. and Tao, T. (2007). The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Ann. Statist.* **35** 2313-2351.
- [12] Chatterjee, S. (2005b). An error bound in the Sudakov-Fernique inequality. arXiv:math/0510424.
- [13] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 257-283.
- [14] Chen, L., and Fang, X. (2011). Multivariate normal approximation by Stein's method: the concentration inequality approach. arXiv:1111.4073.
- [15] Chen, L., Goldstein, L. and Shao, Q.-M. (2011). *Normal Approximation by Stein's Method*. Springer.
- [16] Goldstein, L., and Rinott, Y. (1996) Multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.* **33** 1-17.
- [17] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012a). Gaussian approximation of suprema of empirical processes. arXiv:1212.6906.
- [18] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012b). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. arXiv:1301.4807. Submitted to *Probab. Theory Related Fields*.
- [19] de la Peña, V., Lai, T. and Shao, Q.-M. (2009). *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer.
- [20] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.
- [21] Fan, J., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's  $t$  or bootstrap calibration be applied. *J. Amer. Stat. Assoc.* **102** 1282-1288.
- [22] Frick, K., Marnitz, P. and Munk, A. (2012). Shape-constrained regularization by statistical multiresolution for inverse problems: asymptotic analysis. *Inverse Problems* **28** 065006.
- [23] Gautier, E. and Tsybakov, A. (2011). High-dimensional instrumental variables regression and confidence sets. arXiv: 1105.2454.
- [24] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122-1170.

- [25] Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** 724-739.
- [26] Guerre, E. and Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. *Ann. Statist.* **33** 840-870.
- [27] Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields.* **89** 447-455.
- [28] Horowitz, J. L. and Spokoiny, V.G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69** 599-631.
- [29] Juditsky, A. and A. Nemirovski. (2011). On verifiable sufficient conditions for sparse signal recovery via  $\ell_1$  minimization. *Math. Program. Ser. B* **127** 57-88.
- [30] Koltchinskii, V.I. (1994). Komlós-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoret. Probab.* **7** 73-118.
- [31] Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799-828.
- [32] Komlós, J., Major, P., and Tusnády, G. (1975). An approximation for partial sums of independent rv's and the sample df I. *Z. Warhsch. Verw. Gabiete* **32** 111-131.
- [33] Leadbetter, M., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer.
- [34] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21** 255-285.
- [35] Nagaev, S. (1976). An estimate of the remainder term in the multidimensional central limit theorem. *Proc. Third Japan-USSR Symp. Probab. Theory*. Lecture Notes in Math. pp. 419-438.
- [36] Panchenko, D. (2013). *The Sherrington-Kirkpatrick Model*. Springer.
- [37] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.
- [38] Portnoy, S. (1986). On the central limit theorem in  $\mathbb{R}^p$  when  $p \rightarrow \infty$ . *Probab. Theory Related Fields* **73** 571-583.
- [39] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21-45.
- [40] Romano, J., and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Stat. Assoc.* **100** 94-108.
- [41] Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell Syst. Tech. J.* **41** 463-501.
- [42] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135-1151.
- [43] Talagrand, M. (2003). *Spin Glasses: A Challenge for Mathematicians*. Springer.
- [44] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
- [45] Ye, F. and Zhang, C. (2010). Rate minimaxity of the Lasso and Dantzig selector for the  $\ell_q$  loss in  $\ell_r$  balls. *J. Mach. Learn. Res.* **11** 3519-3540.



# Supplemental Material I for “Central limit theorem and multiplier bootstrap when $p$ is much larger than $n$ ”

## Additional Theoretical Results and Omitted Proofs

V. Chernozhukov, D. Chetverikov, and K. Kato

### APPENDIX E. A NOTE ON RELATION BETWEEN SLEPIAN AND STEIN TYPE METHODS FOR NORMAL APPROXIMATIONS

To keep the notation simple, consider a random vector  $X$  in  $\mathbb{R}^p$  and a standard normal vector  $Z$  in  $\mathbb{R}^p$ . We are interested in bounding

$$\mathbb{E}[g(X)] - \mathbb{E}[g(Z)],$$

over some collection of test functions  $g \in \mathcal{G}$ . Without loss of generality, suppose that  $Z$  and  $X$  are independent.

Consider Stein’s partial differential equation:

$$g(x) - \mathbb{E}[g(Z)] = \Delta h(x) - x' \nabla h(x).$$

It is well known, e.g. [16] and [13], that an explicit solution for  $h$  in this equation is given by

$$h(x) := - \int_0^1 \frac{1}{2t} \left[ \mathbb{E}[g(\sqrt{t}x + \sqrt{1-t}Z)] - \mathbb{E}[g(Z)] \right] dt,$$

so that

$$\mathbb{E}[g(X)] - \mathbb{E}[g(Z)] = \mathbb{E}[\Delta h(X) - X' \nabla h(X)].$$

The Stein type method for normal approximation bounds the right side for  $g \in \mathcal{G}$ .

Next, let us consider the Slepian smart path interpolation:

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Z.$$

Then we have

$$\mathbb{E}[g(X)] - \mathbb{E}[g(Z)] = \mathbb{E} \left[ \int_0^1 \frac{1}{2} \nabla g(Z(t))' \left( \frac{X}{\sqrt{t}} - \frac{Z}{\sqrt{1-t}} \right) dt \right].$$

The Slepian type method, as used in our paper, bounds the right side for  $g \in \mathcal{G}$ .

Elementary calculations and integration by parts yield the following observation.

**Lemma E.1.** *Suppose that  $g : \mathbb{R}^p \rightarrow \mathbb{R}$  is a  $C^2$ -function with uniformly bounded derivatives up to order two. Then*

$$I := \mathbb{E} \left[ \int_0^1 \frac{1}{2} \nabla g(Z(t))' \left( \frac{X}{\sqrt{t}} \right) dt \right] = -\mathbb{E}[X' \nabla h(X)]$$

and

$$II := \mathbb{E} \left[ \int_0^1 \frac{1}{2} \nabla g(Z(t))' \left( \frac{Z}{\sqrt{1-t}} \right) dt \right] = -\mathbb{E}[\Delta h(X)].$$

Hence the Slepian and Stein methods both show that difference between  $I$  and  $II$  is small or approaches zero under suitable conditions on  $X$ ; therefore, they are very similar in spirit, if not identical. The details of treating terms may be different from application to application.

*Proof of Lemma E.1.* By definition of  $h$ , we have

$$-E[X' \nabla h(X)] = E \left[ X' \int_0^1 \frac{1}{2t} \nabla g(Z(t)) \sqrt{t} dt \right] = E \left[ \int_0^1 \nabla g(Z(t))' \frac{X}{2\sqrt{t}} dt \right].$$

On the other hand, by definition of  $h$  and Stein's identity (Lemma E.2),

$$-E[\Delta h(X)] = E \left[ \frac{1}{2} \int_0^1 \Delta g(Z(t)) dt \right] = E \left[ \frac{1}{2} \int_0^1 \nabla g(Z(t))' \left( \frac{Z}{\sqrt{1-t}} \right) dt \right].$$

This completes the proof.  $\blacksquare$

**Lemma E.2** (Stein's identity). *Let  $W = (W_1, \dots, W_p)^T$  be a centered Gaussian random vector in  $\mathbb{R}^p$ . Let  $f : \mathbb{R}^p \rightarrow \mathbb{R}$  be a  $C^1$ -function such that  $E[|\partial_j f(W)|] < \infty$  for all  $1 \leq j \leq p$ . Then for every  $1 \leq j \leq p$ ,*

$$E[W_j f(W)] = \sum_{k=1}^p E[W_j W_k] E[\partial_k f(W)].$$

*Proof of Lemma E.2.* See Section A.6 of [43], and also [42].  $\blacksquare$

## APPENDIX F. OMITTED PROOFS

**F.1. Proof of Lemma A.7.** Claim (a). Define  $I_{ij} = 1\{|x_{ij}| \leq u(\bar{E}[x_{ij}^2])^{1/2}\}$ , and observe that

$$\begin{aligned} (\bar{E}[|\tilde{x}_{ij}|^q])^{1/q} &\leq (\bar{E}[|x_{ij} I_{ij}|^q])^{1/q} + (E_n[|E[x_{ij} I_{ij}]|^q])^{1/q} \\ &\leq (\bar{E}[|x_{ij} I_{ij}|^q])^{1/q} + (\bar{E}[|x_{ij} I_{ij}|^q])^{1/q} \leq 2(\bar{E}[|x_{ij}|^q])^{1/q}. \end{aligned}$$

Claim (b). Observe that

$$\begin{aligned} \bar{E}[|\tilde{x}_{ij} \tilde{x}_{ik} - x_{ij} x_{ik}|] &\leq \bar{E}[|(\tilde{x}_{ij} - x_{ij}) \tilde{x}_{ik}|] + \bar{E}[|x_{ij} (\tilde{x}_{ik} - x_{ik})|] \\ &\leq \sqrt{\bar{E}[(\tilde{x}_{ij} - x_{ij})^2]} \sqrt{\bar{E}[\tilde{x}_{ik}^2]} + \sqrt{\bar{E}[(\tilde{x}_{ik} - x_{ik})^2]} \sqrt{\bar{E}[x_{ij}^2]} \\ &\leq 2\varphi(u) \sqrt{\bar{E}[x_{ij}^2]} \sqrt{\bar{E}[x_{ik}^2]} + \varphi(u) \sqrt{\bar{E}[x_{ik}^2]} \sqrt{\bar{E}[x_{ij}^2]} \\ &\leq (3/2)\varphi(u)(\bar{E}[x_{ij}^2] + \bar{E}[x_{ik}^2]), \end{aligned}$$

where the first inequality follows from the triangle inequality, the second from the Cauchy-Schwarz inequality, the third from the definition of  $\varphi(u)$  together with claim (a), and the last from inequality  $|ab| \leq (a^2 + b^2)/2$ .

Claim (c). This follows from the Cauchy-Schwarz inequality.

Claim (d). We shall use the following lemma.

**Lemma F.1** (Tail Bounds for Self-Normalized Sums). *Let  $\xi_1, \dots, \xi_n$  be independent real-valued random variables such that  $\mathbb{E}[\xi_i] = 0$  and  $\mathbb{E}[\xi_i^2] < \infty$  for all  $1 \leq i \leq n$ . Let  $S_n = \sum_{i=1}^n \xi_i$ . Then for every  $x > 0$ ,*

$$\mathbb{P}(|S_n| > x(4B_n + V_n)) \leq 4\exp(-x^2/2),$$

where  $B_n^2 = \sum_{i=1}^n \mathbb{E}[\xi_i^2]$  and  $V_n^2 = \sum_{i=1}^n \xi_i^2$ .

*Proof of Lemma F.1.* See [19], Theorem 2.16. ■

Define

$$\Lambda_j := 4\sqrt{\bar{\mathbb{E}}[(x_{ij} - \tilde{x}_{ij})^2]} + \sqrt{\mathbb{E}_n[(x_{ij} - \tilde{x}_{ij})^2]}.$$

Then by Lemma F.1 and the union bound, with probability at least  $1 - 4\gamma$ ,

$$|X_j - \tilde{X}_j| \leq \Lambda_j \sqrt{2\log(p/\gamma)}, \text{ for all } 1 \leq j \leq p.$$

By claim (c), for  $u \geq u(\gamma)$ , with probability at least  $1 - \gamma$ , for all  $1 \leq j \leq p$ ,

$$\Lambda_j = 4\sqrt{\bar{\mathbb{E}}[(x_{ij} - \tilde{x}_{ij})^2]} + \sqrt{\mathbb{E}_n[(\mathbb{E}[x_{ij} - \tilde{x}_{ij}])^2]} \leq 5\sqrt{\bar{\mathbb{E}}[x_{ij}^2]}\varphi(u).$$

The last two assertions imply claim (d). ■

**F.2. Proof of Corollary 2.2.** Since  $M_2$  is bounded from below and above by positive constants, we may normalize  $M_2 = 1$ , without loss of generality. In this proof, let  $C > 0$  denote a generic constant depending only on  $c_1$  and  $C_1$ , and its value may change from place to place.

For given  $\gamma \in (0, 1)$ , denote  $\ell_n := \log(pn/\gamma) \geq 1$  and let

$$u_1 := n^{3/8}\ell_n^{-5/8}M_3^{3/4} \text{ and } u_2 := n^{3/8}\ell_n^{-5/8}M_4^{1/2}.$$

Define  $u := u(\gamma) \vee u_1 \vee u_2$  and  $\beta := \sqrt{n}/(2\sqrt{2}u)$ . Then  $u \geq u(\gamma)$  and the choice of  $\beta$  trivially obeys  $2\sqrt{2}u\beta \leq \sqrt{n}$ . So, by Theorem 2.2 and using the argument as that in the proof of Corollary 2.1, for every  $\psi > 0$ , we have for any  $\bar{\varphi}(u) \geq \varphi(u)$ ,

$$(24) \quad \begin{aligned} \rho &\leq C[n^{-1/2}(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3 + (\psi^2 + \psi\beta)\bar{\varphi}(u) \\ &\quad + \psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} + (\beta^{-1}\log p + \psi^{-1})\sqrt{1 \vee \log(p\psi)}]. \end{aligned}$$

**Step 1.** We claim that we can take  $\bar{\varphi}(u) := CM_4^2/u$  for all  $u > 0$ . Since  $\bar{\mathbb{E}}[x_{ij}^2] \geq c_1$ , we have  $1\{|x_{ij}| > u(\bar{\mathbb{E}}[x_{ij}^2])^{1/2}\} \leq 1\{|x_{ij}| > c_1^{1/2}u\}$ . Hence

$$\begin{aligned} \bar{\mathbb{E}}[x_{ij}^2 1\{|x_{ij}| > u(\bar{\mathbb{E}}[x_{ij}^2])^{1/2}\}] &\leq \bar{\mathbb{E}}[x_{ij}^2 1\{|x_{ij}| > c_1^{1/2}u\}] \\ &\leq \bar{\mathbb{E}}[x_{ij}^4 1\{|x_{ij}| > c_1^{1/2}u\}]/(c_1 u^2) \leq \bar{\mathbb{E}}[x_{ij}^4]/(c_1 u^2) \leq M_4^4/(c_1 u^2). \end{aligned}$$

This implies  $\varphi_x(u) \leq CM_4^2/u$ . For  $\varphi_y(u)$ , note that

$$\bar{\mathbb{E}}[y_{ij}^4] = \mathbb{E}_n[\mathbb{E}[y_{ij}^4]] = 3\mathbb{E}_n[(\mathbb{E}[y_{ij}^2])^2] = 3\mathbb{E}_n[(\mathbb{E}[x_{ij}^2])^2] \leq 3\mathbb{E}_n[\mathbb{E}[x_{ij}^4]] = \bar{\mathbb{E}}[x_{ij}^4],$$

and hence  $\varphi_y(u) \leq CM_4^2/u$  as well. This implies the claim of this step.

**Step 2.** We shall bound the right side of (24) by suitably choosing  $\psi$  depending on the range of  $u$ . In order to set up this choice we define  $u^\star$  by the following equation:

$$\bar{\varphi}(u^\star)n^{3/8}/(M_3^3\ell_n^{5/6})^{3/4} = 1.$$

We then take

$$(25) \quad \psi = \psi(u) := \begin{cases} n^{1/8}\ell_n^{-3/8}M_3^{-3/4} & \text{if } u \geq u^\star, \\ \ell_n^{-1/6}(\bar{\varphi}(u))^{-1/3} & \text{if } u < u^\star. \end{cases}$$

We note that for  $u < u^\star$ ,

$$\psi(u) \leq \psi(u^\star) = n^{1/8}\ell_n^{-3/8}M_3^{-3/4}.$$

That is, when  $u < u^\star$  the smoothing parameter  $\psi$  is smaller than when  $u \geq u^\star$ .

Using these choices of parameters  $\beta$  and  $\psi$  and elementary calculations (which will be done in Step 3 below), we conclude from (24) that whether  $u < u^\star$  or  $u \geq u^\star$ ,

$$\rho \leq C(n^{-1/2}u\ell_n^{3/2} + \gamma).$$

The bound in the corollary follows from this inequality.

**Step 3.** (Computation of the bound on  $\rho$ ). Note that since  $\rho \leq 1$ , we only had to consider the case where  $n^{-1/2}u\ell_n^{3/2} \leq 1$  since otherwise the inequality is trivial by taking, say,  $C = 1$ . Since  $u_1 = n^{3/8}M_3^{3/4}/\ell_n^{5/8}$  and  $u_2 = n^{3/8}M_4^{1/2}/\ell_n^{5/8}$ , we have

$$\begin{aligned} (\bar{\varphi}(u^\star))^{4/3} &= n^{-1/2}\ell_n^{5/6}M_3^3, \\ \bar{\varphi}(u_1) &\leq Cn^{-3/8}\ell_n^{5/8}M_4^2/M_3^{3/4}, \\ \bar{\varphi}(u_2) &\leq Cn^{-3/8}\ell_n^{5/8}M_4^{3/2}. \end{aligned}$$

Also note that  $\psi \leq n^{1/8}$ , and so  $1 \vee \log(p\psi) \lesssim \log(pn) \leq \ell_n$ . Therefore,

$$\beta^{-1} \log p \sqrt{1 \vee \log(p\psi)} \lesssim \beta^{-1}\ell_n^{3/2} \lesssim n^{-1/2}u\ell_n^{3/2}.$$

In addition, note that  $\beta \lesssim \sqrt{n}/u \leq \sqrt{n}/u_1 = n^{1/8}\ell_n^{5/8}M_3^{-3/4} =: \bar{\beta}$  and  $\psi \leq \bar{\beta}$  under either case. This implies that  $(\psi^3 + \psi^2\beta + \psi\beta^2) \lesssim \psi\bar{\beta}^2$  and  $(\psi^2 + \psi\beta) \leq \psi\bar{\beta}$ .

Using these inequalities, we can compute the bounds claimed above.

(a). Bounding  $\rho$  when  $u \geq u^\star$ . Then

$$\begin{aligned} n^{-1/2}(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3 &\lesssim n^{-1/2}\psi\bar{\beta}^2M_3^3 \leq n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leq n^{-1/2}u\ell_n^{3/2}; \\ (\psi^2 + \psi\beta)\bar{\varphi}(u) &\lesssim \psi\bar{\beta}\bar{\varphi}(u) \leq \psi\bar{\beta}\bar{\varphi}(u^\star) \leq n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leq n^{-1/2}u\ell_n^{3/2}; \\ \psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} &\leq \psi\bar{\beta}\bar{\varphi}(u)\sqrt{\ell_n/\bar{\beta}} \leq \psi\bar{\beta}\bar{\varphi}(u^\star) \leq n^{-1/2}u\ell_n^{3/2}; \text{ and} \\ \psi^{-1}\sqrt{\ell_n} &\leq n^{-1/8}\ell_n^{7/8}M_3^{3/4} \leq n^{-1/2}u\ell_n^{3/2}; \end{aligned}$$

where we have used Step 1 and the fact that

$$\sqrt{\ell_n}/\bar{\beta} = \ell_n^{-1/2}\psi^{-1} \leq n^{-1/8}\ell_n^{-1/8}M_3^{3/4} \leq n^{-1/2}u\ell_n^{3/2} \leq 1.$$

The claimed bound on  $\rho$  now follows.

**(b).** Bounding  $\rho$  when  $u < u^*$ . Since  $\psi$  is smaller than in case (a), by the calculations in Step (a)

$$n^{-1/2}(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3/\sqrt{n} \lesssim n^{-1/2}u\ell_n^{3/2}.$$

Moreover, using definition of  $\psi$ ,  $u > u_2$ , definition of  $u_2$ , we have

$$\psi\beta\bar{\varphi}(u) \leq \beta\bar{\varphi}(u)^{2/3}\ell_n^{-1/6} \leq \beta\bar{\varphi}(u_2)^{2/3}\ell_n^{-1/6} \leq n^{-1}\beta u_2^2\ell_n^{5/3-1/6} \lesssim n^{-1/2}u\ell_n^{3/2};$$

$$\psi^2\bar{\varphi}(u) \leq \bar{\varphi}(u)^{1/3}\ell_n^{-1/3} \leq \bar{\varphi}(u_2)^{1/3}\ell_n^{-1/3} \leq n^{-1/2}u_2\sqrt{\ell_n} \leq n^{-1/2}u\ell_n^{3/2}.$$

Analogously and using  $n^{-1/2}u\ell_n^{3/2} \leq 1$ , we have

$$\psi\bar{\varphi}(u)\sqrt{\log(p/\gamma)} \leq \bar{\varphi}(u)^{2/3}\ell_n^{1/3} \leq \bar{\varphi}(u_2)^{2/3}\ell_n^{1/3} \leq nu_2^2\ell_n^2 \leq n^{-1/2}u\ell_n^{3/2}.$$

$$\psi^{-1}\sqrt{\ell_n} = \bar{\varphi}(u)^{1/3}\ell_n^{2/3} \leq n^{-1/2}u\ell_n^{3/2}.$$

This completes the proof. ■

## Supplemental Material II for “Central limit theorem and multiplier bootstrap when $p$ is much larger than $n$ ”

### Additional Applications

V. Chernozhukov, D. Chetverikov, and K. Kato

#### APPENDIX G. APPLICATION: MULTIPLE HYPOTHESIS TESTING VIA THE STEPDOWN METHOD

In this section, we study the problem of multiple hypothesis testing in the framework of multiple linear regressions. (Note that the problem of testing multiple means is a special case of testing multiple regressions.) We combine a general stepdown procedure described in [40] with the multiplier bootstrap developed in this paper. In contrast with [40], our results do not require weak convergence arguments, and, thus, can be applied to models with increasing numbers of both parameters and regressions. Notably, the number of regressions can be large in comparison with the sample size.

Let  $(z_i, y_i)_{i=1}^n$  be a sample of independent observations where  $z_i \in \mathbb{R}^p$  is a vector of non-stochastic covariates and  $y_i \in \mathbb{R}^K$  is a vector of dependent random variables. For each  $k = 1, \dots, K$ , let  $I_k \subset \{1, \dots, p\}$  be a subset of covariates used in the  $k$ -th regression. Denote by  $|I_k| = p_k$  the number of covariates in the  $k$ -th regression, and let  $\bar{p} = \max_{1 \leq k \leq K} p_k$ . Let  $v_{ik}$  be a subvector of  $z_i$  consisting of those elements of  $z_i$  whose indices appear in  $I_k$ :  $v_{ik} = (z_{ij})_{j \in I_k}$ . We denote components of  $v_{ik}$  by  $v_{ikj}$ ,  $j = 1, \dots, p_k$ . Without loss of generality, we assume that  $I_k \cap I_{k'} = \emptyset$  for all  $k \neq k'$  and  $\sum_{1 \leq k \leq K} p_k = p$ .

For each  $k = 1, \dots, K$ , consider the linear regression model

$$y_{ik} = v'_{ik} \beta_k + \varepsilon_{ik}, \quad i = 1, \dots, n,$$

where  $\beta_k \in \mathbb{R}^{p_k}$  is an unknown parameter of interest, and  $(\varepsilon_{ik})_{i=1}^n$  is a sequence of independent zero-mean unobservable scalar random variables. We allow for triangular array asymptotics so that everything in the model, and, in particular, the number of regressions  $K$  and the dimensions of the parameters  $\beta_k$  and  $p_k$ , may depend on  $n$ . For brevity, however, we omit index  $n$ . We are interested in simultaneously testing the set of null hypotheses  $H_{kj} : \beta_{kj} = 0$  against the alternatives  $H'_{kj} : \beta_{kj} \neq 0$ ,  $(k, j) \in \mathcal{W}_0$  for some set of pairs  $\mathcal{W}_0$  where  $\beta_{kj}$  denotes the  $j$ th component of  $\beta_k$ , with the strong control of the family-wise error rate. In other words, we seek a procedure that would reject at least one true null hypothesis with probability not greater than  $\alpha + o(1)$  uniformly over the set of true null hypotheses. More formally, let  $\Omega$  be a set of all data generating processes, and  $\omega$  be the true process. Each null hypothesis  $H_{kj}$  is equivalent to  $\omega \in \Omega_{kj}$  for some subset  $\Omega_{kj}$  of  $\Omega$ . Let  $\mathcal{W}$  denote the set of all pairs  $(k, j)$  with  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$ :

$$\mathcal{W} = \{(k, j) : k = 1, \dots, K; j = 1, \dots, p_k\}.$$

For a subset  $w \subset \mathcal{W}$  let  $\Omega^w = (\cap_{(k,j) \in w} \Omega_{kj}) \cap (\cap_{(k,j) \notin w} \Omega_{kj}^c)$  where  $\Omega_{kj}^c = \Omega \setminus \Omega_{kj}$ . The strong control of the family-wise error rate means

$$(26) \quad \sup_{w \subset \mathcal{W}} \sup_{\omega \in \Omega^w} \mathbb{P}\{\text{reject at least one hypothesis among } H_{kj}, (k, j) \in w\} \leq \alpha + o(1).$$

This setting is clearly of interest in many empirical studies.

Our approach is based on the simultaneous analysis of  $t$ -statistics for each component  $\beta_{kj}$ . Let  $x_{ik} = (\mathbb{E}_n[v_{ik}v'_{ik}])^{-1}v_{ik}$ . Then the OLS estimator  $\hat{\beta}_k$  of  $\beta_k$  is given by  $\hat{\beta}_k = \mathbb{E}_n[x_{ik}y_{ik}]$ . The corresponding residuals are  $\hat{\varepsilon}_{ik} = y_{ik} - v'_{ik}\hat{\beta}_k$ ,  $i = 1, \dots, n$ . Since  $(x_{ik})_{i=1}^n$  is non-stochastic, the covariance matrix of  $\hat{\beta}_k$  is given by  $V(\hat{\beta}_k) = \mathbb{E}_n[x_{ik}x'_{ik}\sigma_{ik}^2]/n$  where  $\sigma_{ik}^2 = \mathbb{E}[\varepsilon_{ik}^2]$ ,  $i = 1, \dots, n$ .

The  $t$ -statistic for testing  $H_{kj}$  against  $H'_{kj}$  is  $t_{kj} := |\hat{\beta}_{kj}|/\sqrt{\widehat{V}(\hat{\beta}_k)_{jj}}$  where  $\widehat{V}(\hat{\beta}_k) = \mathbb{E}_n[x_{ik}x'_{ik}\hat{\varepsilon}_{ik}^2]/n$ . Also define

$$t_{kj}^0 := \frac{|\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}^2]}}.$$

Note that  $t_{kj} = t_{kj}^0$  under the hypothesis  $H_{kj}$ .

The stepdown procedure of [40] is described as follows. For a subset  $w \subset \mathcal{W}$ , let  $c_{1-\alpha,w}$  be some estimator of the  $(1-\alpha)$ -quantile of  $\max_{(k,j) \in w} t_{kj}^0$ . On the first step, let  $w(1) = \mathcal{W}_0$ . Reject all hypotheses  $H_{kj}$  satisfying  $t_{kj} > c_{1-\alpha,w(1)}$ . If no null hypothesis is rejected, then stop. If some  $H_{kj}$  are rejected, then let  $w(2)$  be the set of all null hypotheses that were not rejected on the first step. On step  $l \geq 2$ , let  $w(l) \subset \mathcal{W}$  be the subset of null hypotheses that were not rejected up to step  $l$ . Reject all hypotheses  $H_{kj}$ ,  $(k, j) \in w(l)$ , satisfying  $t_{kj} > c_{1-\alpha,w(l)}$ . If no null hypothesis is rejected, then stop. If some  $H_{kj}$  are rejected, then let  $w(l+1)$  be the subset of all null hypotheses among  $(k, j) \in w(l)$  that were not rejected. Proceed in this way until the algorithm stops.

[40] proved the following result. Suppose that  $c_{1-\alpha,w}$  satisfies

$$(27) \quad c_{1-\alpha,w'} \leq c_{1-\alpha,w''} \quad \text{whenever } w' \subset w'',$$

$$(28) \quad \sup_{w \subset \mathcal{W}} \sup_{\omega \in \Omega^w} \mathbb{P}\left(\max_{(k,j) \in w} t_{kj}^0 > c_{1-\alpha,w}\right) \leq \alpha + o(1),$$

then inequality (26) holds. Indeed, let  $w$  be the set of true null hypotheses. Suppose that the procedure rejects at least one of these hypotheses. Let  $l$  be the step when the procedure rejected a true null hypothesis for the first time, and let  $H_{k_0j_0}$  be this hypothesis. Clearly, we have  $w(l) \supset w$ . So,

$$\max_{(k,j) \in w} t_{kj}^0 \geq t_{k_0j_0}^0 = t_{k_0j_0} > c_{1-\alpha,w(l)} \geq c_{1-\alpha,w}.$$

Combining this chain of inequalities with (28) yields (26).

To obtain suitable  $c_{1-\alpha,w}$  that satisfies inequalities (27) and (28) above, we can use the multiplier bootstrap method. Let  $(e_i)_{i=1}^n$  be an i.i.d. sequence

of  $N(0, 1)$  random variables that are independent of the data. Let  $c_{1-\alpha, w}$  be the conditional  $(1 - \alpha)$ -quantile of

$$(29) \quad \max_{(k, j) \in w} \frac{|\sum_{i=1}^n x_{ikj} \hat{\varepsilon}_{ik} e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \hat{\varepsilon}_{ik}^2]}}$$

given  $(z_i, y_i)_{i=1}^n$ . To prove that so defined critical values  $c_{1-\alpha, w}$  satisfy inequalities (27) and (28), we will assume the following regularity condition,

- (M) There are some constants  $c_1 > 0, \bar{\sigma}^2 > 0, \underline{\sigma}^2 > 0$  and a sequence  $B_n \geq 1$  of constants such that for  $1 \leq i \leq n, 1 \leq j \leq p, 1 \leq k \leq K, 1 \leq l \leq p_k$ : (i)  $|z_{ij}| \leq B_n$ ; (ii)  $\mathbb{E}_n[z_{ij}^2] = 1$ ; (iii)  $\underline{\sigma}^2 \leq \mathbb{E}[\varepsilon_{ik}^2] \leq \bar{\sigma}^2$ ; (iv) the minimum eigenvalue of  $\mathbb{E}_n[v_{ik} v'_{ik}]$  is bounded from below by  $c_1$ ; and (v)  $\mathbb{E}_n[x_{ikl}^2] \geq c_1$ .

**Theorem G.1** (Strong Control of Family-Wise Error Rate). *Let  $C_1 > 0$  be some constant and suppose that assumption M is satisfied. Moreover, suppose either*

- (a)  $\mathbb{E}[\max_{1 \leq k \leq K} \varepsilon_{ik}^4] \leq C_1$  for all  $1 \leq i \leq n, \bar{p}^3 B_n^4 (\log p)^4 / n = o(1)$  and  $\bar{p}^2 B_n^4 (\log(pn))^7 / n = o(1)$ ; or
- (b)  $\mathbb{E}[\exp(|\varepsilon_{ik}|/C_1)] \leq 2$  for all  $1 \leq i \leq n, 1 \leq k \leq K, \bar{p}^3 B_n^2 (\log p)^3 / n = o(1)$  and  $\bar{p} B_n^2 (\log(pn))^7 / n = o(1)$ .

*Then the stepdown procedure with the multiplier bootstrap critical values  $c_{1-\alpha, w}$  given above satisfies (26).*

**Comment G.1** (Relation to prior results). There is a vast literature on multiple hypothesis testing. Let us consider the simple case where  $K = p, p_k = 1$  for all  $k = 1, \dots, K$  and  $v_{ik} = 1$ , so that the  $k$ -th regression reduces to  $y_{ik} = \beta_k + \varepsilon_{ik}$  (here  $\beta_k$  is scalar). The problem then reduces to testing multiple means (without stepdown). It is instructive to see the implication of Theorem G.1 in this simple setting. Denote by  $t_k$  the  $t$ -statistic for testing  $H_k : \beta_k = 0$  against  $H'_k : \beta_k \neq 0$ , and let  $c_{1-\alpha}$  be the conditional  $(1 - \alpha)$ -quantile of

$$\max_{k=1, \dots, p} \frac{|\sum_{i=1}^n \hat{\varepsilon}_{ik} e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[\hat{\varepsilon}_{ik}^2]}},$$

where  $\hat{\varepsilon}_{ik} = y_{ik} - \bar{y}_k, \bar{y}_k = \mathbb{E}_n[y_{ik}]$ , and  $(e_i)_{i=1}^n$  is a sequence of i.i.d.  $N(0, 1)$  random variables independent of the data. Theorem G.1 implies that, when  $H_k$  are true for all  $k$ ,  $P(\max_{1 \leq k \leq p} t_k > c_{1-\alpha}) \leq \alpha + o(1)$  (indeed, the inequality “ $\leq$ ” can be replaced by the equality “ $=$ ”) uniformly in the underlying distribution provided that  $\underline{\sigma}^2 \leq \mathbb{E}[\varepsilon_{ik}^2] \leq \bar{\sigma}^2, \log p = o(n^{1/7})$  and either (a)  $\mathbb{E}[\max_{1 \leq k \leq p} \varepsilon_{ik}^4] \leq C_1$  or (b)  $\mathbb{E}[\exp(|\varepsilon_{ik}|/C_1)] \leq 2$ . Hence the multiplier bootstrap as described above leads to an *asymptotically exact* testing procedure for the multiple hypothesis testing problem of which the *logarithm* of the number of hypotheses is nearly of order  $n^{1/7}$  (subject to the prescribed



assumptions). Note here that no assumption on the dependency structure between  $y_{i1}, \dots, y_{ip}$  is made.

The question on how large  $p$  can be was studied in [21] but from a conservative perspective. The motivation there is to know how fast  $p$  can grow to maintain the size of the simultaneous test when we calculate critical values (conservatively) ignoring the dependency among  $t_k$  and assuming that  $t_k$  were distributed as, say,  $N(0, 1)$ . This framework is conservative in that correlation amongst statistics is dealt away with union bounds, namely by Bonferroni-Holm procedures. In contrast, our approach takes into account the correlation amongst statistics and hence is asymptotically exact, that is, asymptotically non-conservative. ■

#### APPENDIX H. APPLICATION: ADAPTIVE SPECIFICATION TESTING

In this section, we study the problem of adaptive specification testing. Let  $(v_i, y_i)_{i=1}^n$  be a sample of independent random pairs where  $y_i$  is a scalar dependent random variable, and  $v_i \in \mathbb{R}^d$  is a vector of non-stochastic covariates. The null hypothesis,  $H_0$ , is that there exists  $\beta \in \mathbb{R}^d$  such that

$$(30) \quad \mathbb{E}[y_i] = v_i' \beta; \quad i = 1, \dots, n.$$

The alternative hypothesis,  $H_a$ , is that there is no  $\beta$  satisfying (30). We allow for triangular array asymptotics so that everything in the model may depend on  $n$ . For brevity, however, we omit index  $n$ .

Let  $\varepsilon_i = y_i - \mathbb{E}[y_i]$ ,  $i = 1, \dots, n$ . Then  $\mathbb{E}[\varepsilon_i] = 0$ , and under  $H_0$ ,  $y_i = v_i' \beta + \varepsilon_i$ . To test  $H_0$ , consider a set of test functions  $P_j(v_i)$ ,  $j = 1, \dots, p$ . Let  $z_{ij} = P_j(v_i)$ . We choose test functions so that  $\mathbb{E}_n[z_{ij} v_i] = 0$  and  $\mathbb{E}_n[z_{ij}^2] = 1$  for all  $j = 1, \dots, p$ . In our analysis,  $p$  may be higher or even much higher than  $n$ . Let  $\hat{\beta} = (\mathbb{E}_n[v_i v_i'])^{-1} (\mathbb{E}_n[v_i y_i])$  be an OLS estimator of  $\beta$ , and let  $\hat{\varepsilon}_i = y_i - z_i' \hat{\beta}$ ;  $i = 1, \dots, n$  be corresponding residuals. Our test statistic is

$$T := \max_{1 \leq j \leq p} \frac{|\sum_{i=1}^n z_{ij} \hat{\varepsilon}_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2 \hat{\varepsilon}_i^2]}}.$$

The test rejects  $H_0$  if  $T$  is significantly large.

Note that since  $\mathbb{E}_n[z_{ij} v_i] = 0$ , we have

$$\sum_{i=1}^n z_{ij} \hat{\varepsilon}_i / \sqrt{n} = \sum_{i=1}^n z_{ij} (\varepsilon_i + v_i' (\beta - \hat{\beta})) / \sqrt{n} = \sum_{i=1}^n z_{ij} \varepsilon_i / \sqrt{n}.$$

Therefore, under  $H_0$ ,

$$T = \max_{1 \leq j \leq p} \frac{|\sum_{i=1}^n z_{ij} \varepsilon_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2 \varepsilon_i^2]}}.$$

This suggests that we can use the multiplier bootstrap to obtain a critical value for the test. More precisely, let  $(e_i)_{i=1}^n$  be a sequence of independent

$N(0, 1)$  random variables that are independent of the data, and let

$$W := \max_{1 \leq j \leq p} \frac{|\sum_{i=1}^n z_{ij} \widehat{\varepsilon}_i e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2 \widehat{\varepsilon}_i^2]}}.$$

The multiplier bootstrap critical value  $c_W(1 - \alpha)$  is the conditional  $(1 - \alpha)$ -quantile of  $W$  given the data. To prove the validity of multiplier bootstrap, we will impose the following condition:

- (S) There are some constants  $c_1 > 0, C_1 > 0, \bar{\sigma}^2 > 0, \underline{\sigma}^2 > 0$ , and a sequence  $B_n \geq 1$  of constants such that for all  $1 \leq i \leq n, 1 \leq j \leq p, 1 \leq k \leq d$ : (i)  $|z_{ij}| \leq B_n$ ; (ii)  $\mathbb{E}_n[z_{ij}^2] = 1$ ; (iii)  $\underline{\sigma}^2 \leq \mathbb{E}[\varepsilon_i^2] \leq \bar{\sigma}^2$ ; (iv)  $|v_{ik}| \leq C_1$ ; (v)  $d \leq C_1$ ; and (vi) the minimum eigenvalue of  $\mathbb{E}_n[v_i v_i']$  is bounded from below by  $c_1$ .

**Theorem H.1** (Size Control of Adaptive Specification Test). *Let  $c_2 > 0$  be some constant. Suppose that assumption S is satisfied. Moreover, suppose that either*

- (a)  $\mathbb{E}[\varepsilon_i^4] \leq C_1$  for all  $1 \leq i \leq n$  and  $B_n^4(\log(pn))^7/n \leq C_1 n^{-c_2}$ ; or  
 (b)  $\mathbb{E}[\exp(|\varepsilon_i|/C_1)] \leq 2$  for all  $1 \leq i \leq n$  and  $B_n^2(\log(pn))^7/n \leq C_1 n^{-c_2}$ .

*Then there exist constants  $c > 0$  and  $C > 0$ , depending only on  $c_1, c_2, C_1, \underline{\sigma}^2$  and  $\bar{\sigma}^2$ , such that under  $H_0$ ,  $|\mathbb{P}(T \leq c_W(1 - \alpha)) - (1 - \alpha)| \leq C n^{-c}$ .*

**Comment H.1.** The literature on specification testing is large. In particular, [28] and [26] developed adaptive tests that are suitable for inference in  $L_2$ -norm. In contrast, our test is most suitable for inference in sup-norm. An advantage of our procedure is that selecting a wide class of test functions leads to a test that can effectively adapt to a wide range of alternatives, including those that can not be well-approximated by Hölder-continuous functions. ■

## APPENDIX I. PROOFS FOR SECTION G

**I.1. Proof of Theorem G.1.** The multiplier bootstrap critical value  $c_{1-\alpha, w}$  clearly satisfies  $c_{1-\alpha, w} \leq c_{1-\alpha, w'}$  whenever  $w \subset w'$ , so inequality (27) is satisfied. Therefore, it suffices to prove (28). For the notational convenience, we will only consider the  $w = \mathcal{W}$  case and suppress the uniformity in the underlying distribution. The general case follows from inspection of the proof.

Let us define

$$T := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj} \varepsilon_{ik} / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik}^2]}}, \quad W := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj} \widehat{\varepsilon}_{ik} e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \widehat{\varepsilon}_{ik}^2]}}.$$

We shall prove that  $\mathbb{P}(T > c_W(1 - \alpha)) = \alpha + o(1)$ , where recall that  $c_W(1 - \alpha)$  is the conditional  $(1 - \alpha)$ -quantile of  $W$  given  $(\varepsilon_{ik})$ . Here we will only consider case (a) of the theorem. The proof for case (b) is similar and hence omitted.

We make use of Corollary 3.1-(iv) to prove the desired claim. Define

$$T_0 := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj} \varepsilon_{ik} / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \sigma_i^2]}}, \quad W_0 := \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj} \varepsilon_{ik} e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \sigma_i^2]}}.$$

We first verify conditions (14) and (15) in Section 3. We will use the following facts directly deduced from assumption M:

$$(31) \quad \begin{aligned} \max_{i,k,j} |x_{ikj}| &\leq \max_{i,k} \|x_{ik}\| \leq_{(1)} c_1^{-1} \max_{i,k} \|v_{ik}\| \\ &\leq c_1^{-1} \sqrt{\bar{p}} \max_{i,k,j} |v_{ikj}| \leq_{(2)} c_1^{-1} \sqrt{\bar{p}} B_n, \end{aligned}$$

$$(32) \quad \begin{aligned} \max_{k,j} \mathbb{E}_n[x_{ikj}^2] &\leq \max_k \mathbb{E}_n[\|x_{ik}\|^2] \\ &\leq_{(3)} c_1^{-2} \max_k \mathbb{E}_n[\|v_{ik}\|^2] \leq_{(4)} c_1^{-2} \bar{p}, \end{aligned}$$

where (1) and (3) follow from assumption M-(iv) and definition of  $x_{ik}$ , (2) is from M-(i) since  $v_{ik}$  is a subvector of  $z_i$ , and (4) is due to M-(ii). We shall first prove some lemmas. In these lemmas, we will assume all the conditions in Theorem G.1 case (a) without mentioning so.

**Lemma I.1.**  $\sum_{i=1}^n x_{ikj} \varepsilon_{ik} / \sqrt{n} = O_P(r_{n1})$  uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$  where  $r_{n1} = \sqrt{\bar{p} \log p}$ .

*Proof.* By Lemma A.1 combined with inequalities (31) and (32), we have  $\mathbb{E}[\max_{k,j} |\sum_{i=1}^n x_{ikj} \varepsilon_{ik} / \sqrt{n}|] = O(\sqrt{\bar{p}} B_n (\log p) / n^{1/4} + \sqrt{\bar{p} \log p}) = O(\sqrt{\bar{p} \log p})$ ,

where the second step follows because  $B_n \sqrt{\log p} / n^{1/4} = o(1)$ . The claim follows from Markov's inequality.  $\blacksquare$

**Lemma I.2.**  $\mathbb{E}_n[x_{ikj}^2 (\hat{\varepsilon}_{ik}^2 - \sigma_{ik}^2)] = O_P(r_{n2})$  uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$  where  $r_{n2} = \bar{p} B_n^2 (\log p) / \sqrt{n}$ .

*Proof.* We have

$$\begin{aligned} \mathbb{E}_n[x_{ikj}^2 (\hat{\varepsilon}_{ik}^2 - \sigma_{ik}^2)] &= \mathbb{E}_n[x_{ikj}^2 (\varepsilon_{ik}^2 - \sigma_{ik}^2)] + \mathbb{E}_n[x_{ikj}^2 (v'_{ik} (\hat{\beta}_k - \beta_k))^2] \\ &\quad - 2\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} v'_{ik} (\hat{\beta}_k - \beta_k)] \\ &=: I_{jk} + II_{jk} + III_{jk}. \end{aligned}$$

We will show in steps 1-3 below that  $I_{jk} = O_P(\bar{p} B_n^2 (\log p) / \sqrt{n})$ ,  $II_{jk} = O_P(\bar{p}^2 B_n^2 (\log p) / n)$ , and  $III_{jk} = O_P(\bar{p}^2 B_n^2 (\log p) / n)$  uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$ . The claim of the lemma follows since  $\bar{p} / \sqrt{n} \rightarrow 0$ .

**Step 1.** We prove that  $I_{jk} = \mathbb{E}_n[x_{ikj}^2 (\varepsilon_{ik}^2 - \sigma_{ik}^2)] = O_P(\bar{p} B_n^2 (\log p) / \sqrt{n})$  uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$ .

By Lemma A.1 combined with inequalities (31) and (32), we have

$$\begin{aligned} \mathbb{E}[\max_{k,j} |\mathbb{E}_n[x_{ikj}^2 (\varepsilon_{ik}^2 - \sigma_{ik}^2)]|] &= O(\bar{p} B_n^2 (\log p) / \sqrt{n} + \bar{p} B_n \sqrt{(\log p) / n}) \\ &= O(\bar{p} B_n^2 (\log p) / \sqrt{n}), \end{aligned}$$

where the second step follows because  $B_n \geq 1$ . The claim of this step follows from Markov's inequality.

**Step 2.** We prove that

$$II_{jk} = \mathbb{E}_n[x_{ikj}^2(v'_{ik}(\hat{\beta}_k - \beta_k))^2] = O_P(\bar{p}^2 B_n^2 (\log p)/n)$$

uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$ . We have

$$\begin{aligned} \max_{k,j} \mathbb{E}_n[x_{ikj}^2(v'_{ik}(\hat{\beta}_k - \beta_k))^2] &\leq_{(1)} c_1^{-2} \bar{p} B_n^2 \max_k \mathbb{E}_n[(v'_{ik}(\hat{\beta}_k - \beta_k))^2] \\ &= c_1^{-2} \bar{p} B_n^2 \max_k \mathbb{E}_n[\varepsilon_{ik} v'_{ik}] \mathbb{E}_n[v_{ik} v'_{ik}]^{-1} \mathbb{E}_n[v_{ik} \varepsilon_{ik}] \\ &\leq_{(2)} c_1^{-3} \bar{p} B_n^2 \max_k \|\mathbb{E}_n[v_{ik} \varepsilon_{ik}]\|^2 \\ &\leq c_1^{-3} \bar{p}^2 B_n^2 \max_{k,j} |\mathbb{E}_n[v_{ikj} \varepsilon_{ik}]|^2 \\ &=_{(3)} O_P(\bar{p}^2 B_n^2 (\log p)/n), \end{aligned}$$

where (1) follows from inequality (31), (2) from assumption M-(iv), and (3) from application of Lemma A.1. The claim of this step follows.

**Step 3.** We prove that

$$III_{jk} = \mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} (v'_{ik}(\hat{\beta}_k - \beta_k))] = O_P(\bar{p}^2 B_n^2 (\log p)/n)$$

uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$ . We have

$$\begin{aligned} \max_{k,j} |\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} (v'_{ik}(\hat{\beta}_k - \beta_k))]| &\leq \max_{k,j} \|\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} v'_{ik}]\| \|\hat{\beta}_k - \beta_k\| \\ &\leq \max_{k,j,l} \sqrt{\bar{p}} |\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} v_{ikl}]| \|\hat{\beta}_k - \beta_k\|. \end{aligned}$$

Then

$$\begin{aligned} \max_k \|\hat{\beta}_k - \beta_k\| &= \max_k \|\mathbb{E}_n[v_{ik} v'_{ik}]^{-1} \mathbb{E}_n[v_{ik} \varepsilon_{ik}]\| \leq_{(1)} c_1^{-1} \max_k \|\mathbb{E}_n[v_{ik} \varepsilon_{ik}]\| \\ &\leq c_1^{-1} \sqrt{\bar{p}} \max_{k,j} |\mathbb{E}_n[v_{ikj} \varepsilon_{ik}]| =_{(2)} O_P(\sqrt{\bar{p} (\log p)/n}) \end{aligned}$$

where (1) follows from assumption M-(iv) and (2) is as in step 2. In addition, by Lemma A.1 combined with inequalities (31) and (32), we have

$$\begin{aligned} \mathbb{E}[\max_{k,j,l} |\mathbb{E}_n[x_{ikj}^2 \varepsilon_{ik} v_{ikl}]|] &= O(\bar{p} B_n^3 (\log p)/n^{3/4} + \bar{p} B_n^2 \sqrt{(\log p)/n}) \\ &= O(\bar{p} B_n^2 \sqrt{(\log p)/n}), \end{aligned}$$

where the last step is because  $B_n \sqrt{\log p}/n^{1/4} = o(1)$ . Combining these bounds yields the claim of this step.  $\blacksquare$

In Lemmas I.3 and I.4,  $\mathbb{E}_e[\cdot]$  denotes the expectation with respect to  $(e_i)_{i=1}^n$  conditional on  $(\varepsilon_{ik})$ .

**Lemma I.3.**  $\sum_{i=1}^n x_{ikj} \hat{\varepsilon}_{ik} e_i / \sqrt{n} = O_P(r_{n1})$  uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$ . Recall that  $r_{n1} = \sqrt{\bar{p} \log p}$ .

*Proof.* We have

$$\begin{aligned}
\mathbb{E}_e[\max_{k,j} |\sum_{i=1}^n x_{ikj} \hat{\varepsilon}_{ik} e_i / \sqrt{n}|] &\lesssim_{(1)} \sqrt{\log p} \max_{k,j} (\mathbb{E}_n[x_{ikj}^2 \hat{\varepsilon}_{ik}^2])^{1/2} \\
&=_{(2)} \sqrt{\log p} \max_{k,j} (\mathbb{E}_n[x_{ikj}^2 \sigma_{ik}^2] + O_P(r_{n2}))^{1/2} \\
&\leq_{(3)} \sqrt{\log p} \max_{k,j} (\mathbb{E}_n[x_{ikj}^2 \sigma_{ik}^2])^{1/2} + O_P(r_{n2} \sqrt{\log p}) \\
&\leq_{(4)} \sigma \sqrt{\log p} \max_{k,j} (\mathbb{E}_n[x_{ikj}^2])^{1/2} + O_P(r_{n2} \sqrt{\log p}) \\
&=_{(5)} O_P(\sqrt{\bar{p} \log p}),
\end{aligned}$$

where (1) follows from Pisier's inequality, (2) from lemma I.2, (3) follows from application of Taylor's theorem together with the fact that  $r_{n2} = o(1)$  and  $\mathbb{E}_n[x_{ikj}^2 \sigma_{ik}^2]$  is bounded away from zero (which is guaranteed by assumptions M-(iii) and M-(v)) (4) follows from assumption M-(iii), and (5) is due to equation (32) and  $r_{n2} = o(1)$ . The claim of the lemma follows. ■

**Lemma I.4.**  $\sum_{i=1}^n x_{ikj}(\hat{\varepsilon}_{ik} - \varepsilon_{ik})e_i / \sqrt{n} = O_P(r_{n3})$  uniformly over  $k = 1, \dots, K$  and  $j = 1, \dots, p_k$  where  $r_{n3} = \bar{p} B_n(\log p) / \sqrt{n}$ .

*Proof.* We have

$$\begin{aligned}
\mathbb{E}_e[|\sum_{i=1}^n x_{ikj}(\hat{\varepsilon}_{ik} - \varepsilon_{ik})e_i / \sqrt{n}|] &\lesssim_{(1)} \sqrt{\log p} \max_{k,j} (\mathbb{E}_n[x_{ikj}^2 (\hat{\varepsilon}_{ik} - \varepsilon_{ik})^2])^{1/2} \\
&=_{(2)} \sqrt{\log p} \max_{k,j} (\mathbb{E}_n[x_{ikj}^2 (v'_{ik}(\hat{\beta}_k - \beta_k))^2])^{1/2} \\
&=_{(3)} O_P(\bar{p} B_n(\log p) / \sqrt{n})
\end{aligned}$$

where (1) follows from Pisier's inequality, (2) is by definition of  $\hat{\varepsilon}_{ik}$ , and (3) is by step 2 in the proof of lemma I.2. The claim follows. ■

Going back to the proof of Theorem G.1, by Lemmas I.1 and I.2 and the fact that  $\mathbb{E}_n[x_{ikj}^2 \sigma_{ik}^2]$  is bounded away from zero, we have

$$\begin{aligned}
T &= \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj} \varepsilon_{ik} / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \sigma_{ik}^2] + O_P(r_{n2})}} \\
&= T_0 + O_P(r_{n1} r_{n2}) = T_0 + o_P(1 / \sqrt{\log p}),
\end{aligned}$$

where the last step uses the fact that  $\bar{p}^3 B_n^4(\log p)^4 / n = o(1)$ . Similarly, by Lemmas I.2-I.4, we have

$$\begin{aligned}
W &= \max_{k,j} \frac{|\sum_{i=1}^n x_{ikj} \hat{\varepsilon}_{ik} e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2 \sigma_{ik}^2]}} + O_P(r_{n1} r_{n2}) \\
&= W_0 + O_P(r_{n1} r_{n2} + r_{n3}) = W_0 + o_P(1 / \sqrt{\log p}),
\end{aligned}$$

where the last step uses the fact that  $\bar{p} B_n(\log p)^{3/2} / \sqrt{n} = o(1)$ . Hence it is verified that conditions (14) and (15) in Section 3 are satisfied with some

sequences  $\zeta_1 = \zeta_{1n} \rightarrow 0$  and  $\zeta_2 = \zeta_{2n} \rightarrow 0$  such that  $\zeta_1 \sqrt{\log p} + \zeta_2 = o(1)$ . Therefore, the desired claim follows from Corollary 3.1-(iv).  $\blacksquare$

## APPENDIX J. PROOFS FOR SECTION H

**J.1. Proof of Theorem H.1.** We only consider case (a). The proof for case (b) is similar and hence omitted. In this proof, let  $c, c', C, C'$  denote generic positive constants depending only on  $c_1, c_2, C_1, \underline{\sigma}^2, \bar{\sigma}^2$  and their values may change from place to place. Let

$$T_0 := \max_{1 \leq j \leq p} \frac{|\sum_{i=1}^n z_{ij} \varepsilon_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2 \sigma_i^2]}} \text{ and } W_0 := \max_{1 \leq j \leq p} \frac{|\sum_{i=1}^n z_{ij} \varepsilon_i e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2 \sigma_i^2]}}.$$

We make use of Corollary 3.1-(iv). To this end, we shall verify conditions (14) and (15) in Section 3, which will be separately done in Steps 1 and 2, respectively.

**Step 1.** We show that  $P(|T - T_0| > \zeta_1) < \zeta_2$  for some  $\zeta_1$  and  $\zeta_2$  satisfying  $\zeta_1 \sqrt{\log p} + \zeta_2 \leq Cn^{-c}$ .

By Corollary 2.3-(v), we have

$$\begin{aligned} & P\left(\max_{1 \leq j \leq p} |\sum_{i=1}^n z_{ij} \varepsilon_i / \sqrt{n}| > t\right) \\ & \leq P\left(\max_{1 \leq j \leq p} |\sum_{i=1}^n z_{ij} \sigma_i e_i / \sqrt{n}| > t\right) + Cn^{-c}, \end{aligned}$$

uniformly in  $t \in \mathbb{R}$ . By the Gaussian Concentration Inequality, for every  $t > 0$ , we have

$$P\left(\max_{1 \leq j \leq p} |\sum_{i=1}^n z_{ij} \sigma_i e_i / \sqrt{n}| > \mathbb{E}[\max_{1 \leq j \leq p} |\sum_{i=1}^n z_{ij} \sigma_i e_i / \sqrt{n}|] + Ct\right) \leq e^{-t^2}.$$

Since  $\mathbb{E}[\max_{1 \leq j \leq p} |\sum_{i=1}^n z_{ij} \sigma_i e_i / \sqrt{n}|] \leq C\sqrt{\log p}$ , we conclude that

$$(33) \quad P\left(\max_{1 \leq j \leq p} |\sum_{i=1}^n z_{ij} \varepsilon_i / \sqrt{n}| > C\sqrt{\log(pn)}\right) \leq C'n^{-c}.$$

Moreover,

$$\begin{aligned} \mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i^2 - \sigma_i^2)] &= \mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i - \varepsilon_i)^2] + \mathbb{E}_n[z_{ij}^2(\varepsilon_i^2 - \sigma_i^2)] + 2\mathbb{E}_n[z_{ij}^2 \varepsilon_i(\hat{\varepsilon}_i - \varepsilon_i)] \\ &=: I_j + II_j + III_j. \end{aligned}$$

Consider  $I_j$ . We have

$$I_j \leq_{(1)} \max_{1 \leq i \leq n} (\hat{\varepsilon}_i - \varepsilon_i)^2 \leq_{(2)} C\|\hat{\beta} - \beta\|^2 \leq_{(3)} C'\|\mathbb{E}_n[v_i \varepsilon_i]\|^2,$$

where (1) follows from assumption S-(ii), (2) from S-(iv) and S-(v), and (3) from S-(vi). Since  $\mathbb{E}[\|\mathbb{E}_n[v_i \varepsilon_i]\|^2] \leq C/n$ , by Markov's inequality, for every  $t > 0$ ,

$$(34) \quad P\left(\max_{1 \leq j \leq p} \mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i - \varepsilon_i)^2] > t\right) \leq C/(nt).$$

Consider  $II_j$ . By Lemma A.1 and Markov's inequality, we have

$$(35) \quad \mathbb{P} \left( \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2(\varepsilon_i^2 - \sigma_i^2)]| > t \right) \leq CB_n^2(\log p)/(\sqrt{nt}).$$

Consider  $III_j$ . We have  $|III_j| \leq 2|\mathbb{E}_n[z_{ij}^2 v_i'(\beta - \hat{\beta})\varepsilon_i]| \leq 2\|\mathbb{E}_n[z_{ij}^2 \varepsilon_i v_i]\| \|\hat{\beta} - \beta\|$ . Hence

$$(36) \quad \begin{aligned} & \mathbb{P} \left( \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2 \varepsilon_i(\hat{\varepsilon}_i - \varepsilon_i)]| > t \right) \\ & \leq \mathbb{P} \left( \max_{1 \leq j \leq p} \|\mathbb{E}_n[z_{ij}^2 \varepsilon_i v_i]\| > t \right) + \mathbb{P}(\|\hat{\beta} - \beta\| > 1) \\ & \leq C[B_n^2(\log p)/(\sqrt{nt}) + 1/n]. \end{aligned}$$

By (34)-(36), we have

$$(37) \quad \mathbb{P} \left( \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i^2 - \sigma_i^2)]| > t \right) \leq C[B_n^2(\log p)/(\sqrt{nt}) + 1/(nt) + 1/n].$$

In particular,

$$\mathbb{P} \left( \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i^2 - \sigma_i^2)]| > \underline{\sigma}^2/2 \right) \leq Cn^{-c}.$$

Since  $\mathbb{E}_n[z_{ij}^2 \sigma_i^2] \geq \underline{\sigma}^2 > 0$  (which is guaranteed by S-(iii) and S-(ii)), on the event  $\max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i^2 - \sigma_i^2)]| \leq \underline{\sigma}^2/2$ , we have

$$\min_{1 \leq j \leq p} \mathbb{E}_n[z_{ij}^2 \hat{\varepsilon}_i^2] \geq \min_{1 \leq j \leq p} \mathbb{E}_n[z_{ij}^2 \sigma_i^2] - \underline{\sigma}^2/2 \geq \underline{\sigma}^2/2,$$

and hence

$$\begin{aligned} |T - T_0| &= \max_{1 \leq j \leq p} \left| \frac{\sqrt{\mathbb{E}_n[z_{ij}^2 \sigma_i^2]} - \sqrt{\mathbb{E}_n[z_{ij}^2 \hat{\varepsilon}_i^2]}}{\sqrt{\mathbb{E}_n[z_{ij}^2 \hat{\varepsilon}_i^2]}} \right| \times T_0 \\ &\leq C \max_{1 \leq j \leq p} \left| \sqrt{\mathbb{E}_n[z_{ij}^2 \sigma_i^2]} - \sqrt{\mathbb{E}_n[z_{ij}^2 \hat{\varepsilon}_i^2]} \right| \times T_0 \\ &\leq C \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2 \sigma_i^2] - \mathbb{E}_n[z_{ij}^2 \hat{\varepsilon}_i^2]| \times T_0, \end{aligned}$$

where the last step uses the simple fact that

$$|\sqrt{a} - \sqrt{b}| = \frac{|a - b|}{\sqrt{a} + \sqrt{b}} \leq \frac{|a - b|}{\sqrt{a}}.$$

By (33) and (37), for every  $t > 0$ ,

$$\mathbb{P} \left( |T - T_0| > Ct\sqrt{\log(pn)} \right) \leq C'[n^{-c} + B_n^2(\log p)/(\sqrt{nt}) + 1/(nt)].$$

By choosing  $t = (\log(pn))^{-1}n^{-c'}$  with sufficiently small  $c' > 0$ , we obtain the claim of this step.

**Step 2.** We show that  $\mathbb{P}(\mathbb{P}_e(|W - W_0| > \zeta_1) > \zeta_2) < \zeta_2$  for some  $\zeta_1$  and  $\zeta_2$  satisfying  $\zeta_1\sqrt{\log p} + \zeta_2 \leq Cn^{-c}$ .

For  $0 < t \leq \underline{\sigma}^2/2$ , consider the event

$$\mathcal{E} = \left\{ (\varepsilon_i)_{i=1}^n : \max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2(\hat{\varepsilon}_i^2 - \sigma_i^2)]| \leq t, \max_{1 \leq i \leq p} (\hat{\varepsilon}_i - \varepsilon_i)^2 \leq t^2 \right\}.$$

By calculations in Step 1,  $P(\mathcal{E}) \geq 1 - C[B_n^2(\log p)/(\sqrt{nt}) + 1/(nt^2) + 1/n]$ . We shall show that, on this event,

$$(38) \quad P_e \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^n z_{ij} \hat{\varepsilon}_i e_i / \sqrt{n} \right| > C \sqrt{\log(pn)} \right) \leq n^{-1},$$

$$(39) \quad P_e \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^n z_{ij} (\hat{\varepsilon}_i - \varepsilon_i) e_i / \sqrt{n} \right| > Ct \sqrt{\log(pn)} \right) \leq n^{-1}.$$

For (38), by the Gaussian concentration inequality, for every  $s > 0$ ,

$$P_e \left( \max_{1 \leq j \leq p} \left| \sum_{i=1}^n z_{ij} \hat{\varepsilon}_i e_i / \sqrt{n} \right| > \mathbb{E}_e \left[ \max_{1 \leq j \leq p} \left| \sum_{i=1}^n z_{ij} \hat{\varepsilon}_i e_i / \sqrt{n} \right| \right] + Cs \right) \leq e^{-s^2}.$$

where we have used the fact  $\mathbb{E}_n[z_{ij}^2 \hat{\varepsilon}_i^2] = \mathbb{E}_n[z_{ij}^2 \sigma_i^2] + \mathbb{E}_n[z_{ij}^2 (\hat{\varepsilon}_i^2 - \sigma_i^2)] \leq \bar{\sigma}^2 + t \leq \bar{\sigma}^2 + \underline{\sigma}^2/2$  on the event  $\mathcal{E}$ . Here  $\mathbb{E}_e[\cdot]$  means the expectation with respect to  $(e_i)_{i=1}^n$  conditional on  $(\varepsilon_i)_{i=1}^n$ . Moreover, on the event  $\mathcal{E}$ ,

$$\mathbb{E}_e \left[ \max_{1 \leq j \leq p} \left| \sum_{i=1}^n z_{ij} \hat{\varepsilon}_i e_i / \sqrt{n} \right| \right] \leq C \sqrt{\log p}.$$

Hence by choosing  $s = \sqrt{\log n}$ , we obtain (38). Inequality (39) follows similarly, by noting that  $(\mathbb{E}_n[z_{ij}^2 (\hat{\varepsilon}_i - \varepsilon_i)^2])^{1/2} \leq \max_{1 \leq i \leq n} |\hat{\varepsilon}_i - \varepsilon_i| \leq t$  on the event  $\mathcal{E}$ .

Define

$$W_1 := \max_{1 \leq j \leq p} \frac{\left| \sum_{i=1}^n z_{ij} \hat{\varepsilon}_i e_i / \sqrt{n} \right|}{\sqrt{\mathbb{E}_n[z_{ij}^2 \sigma_i^2]}}.$$

Note that  $\mathbb{E}_n[z_{ij}^2 \sigma_i^2] \geq \underline{\sigma}^2$ . Since on the event  $\mathcal{E}$ ,  $\max_{1 \leq j \leq p} |\mathbb{E}_n[z_{ij}^2 (\hat{\varepsilon}_i^2 - \sigma_i^2)]| \leq t \leq \underline{\sigma}^2/2$ , in view of Step 1, on this event, we have

$$\begin{aligned} |W - W_0| &\leq |W - W_1| + |W_1 - W_0| \\ &\leq CtW_1 + |W_1 - W_0| \\ &\leq Ct \max_{1 \leq j \leq p} \left| \sum_{i=1}^n z_{ij} \hat{\varepsilon}_i e_i / \sqrt{n} \right| + C \max_{1 \leq j \leq p} \left| \sum_{i=1}^n z_{ij} (\hat{\varepsilon}_i - \varepsilon_i) e_i / \sqrt{n} \right|. \end{aligned}$$

Therefore, by (38) and (39), on the event  $\mathcal{E}$ , we have

$$P_e \left( |W - W_0| > Ct \sqrt{\log(pn)} \right) \leq 2n^{-1}.$$

By choosing  $t = (\log(pn))^{-1} n^{-c}$  with sufficiently small  $c > 0$ , we obtain the claim of this step.

**Step 3.** Steps 1 and 2 verified conditions (14) and (15) in Section 3. Theorem H.1 case (a) follows from Corollary 3.1-(iv).  $\blacksquare$

(V. Chernozhukov) DEPARTMENT OF ECONOMICS AND OPERATIONS RESEARCH CENTER, MIT, 50 MEMORIAL DRIVE, CAMBRIDGE, MA 02142, USA.

*E-mail address:* vchern@mit.edu



(D. Chetverikov) DEPARTMENT OF ECONOMICS, MIT, 50 MEMORIAL DRIVE, CAMBRIDGE, MA 02142, USA.

*E-mail address:* `dchetver@mit.edu`

(K. Kato) DEPARTMENT OF MATHEMATICS, GRADUATE SCHOOL OF SCIENCE, HIROSHIMA UNIVERSITY, 1-3-1 KAGAMIYAMA, HIGASHI-HIROSHIMA, HIROSHIMA 739-8526, JAPAN.

*E-mail address:* `kkato@hiroshima-u.ac.jp`